

ББК 32.973.26-018,2.75

М62

УДК 681.3.07

Компьютерное изд-во "Диалектика"

Зав. редакцией *А.В. Слепцов*

По общим вопросам обращайтесь в издательство "Диалектика" по адресу:
info@dialektika.com, <http://www.dialektika.com>

Минько, А.А.

М62 Статистический анализ в MS Excel. : — М. : Издательский дом "Вильямс", 2004. — 448 с. : ил. — Парал. тит. англ.

ISBN 5-8459-0692-X (рус.)

Книга предназначена для всех, кто использует методы статистического анализа в своей работе. Она написана как "сборник рецептов" статистических методов, которые часто применяются на практике и которые сравнительно просто реализуются в электронной таблице Excel. Для каждого приведенного метода четко описана статистическая модель, в рамках которой его можно применять. Кроме того, методы сгруппированы по типу исходных данных, предъявляемых для статистического анализа. Методы представлены в таком виде, чтобы их могли легко отобрать для своих потребностей и сравнительно просто реализовать практические работники, которым необходимо самостоятельно провести статистический анализ своих данных.

Для студентов, аспирантов, преподавателей и практических работников, занимающихся вопросами анализа и обработки статистических данных.

ББК 32.973.26-018.2.75

Все названия программных продуктов являются зарегистрированными торговыми марками соответствующих фирм.

Никакая часть настоящего издания ни в каких целях не может быть воспроизведена в какой бы то ни было форме и какими бы то ни было средствами, будь то электронные или механические, включая фотокопирование и запись на магнитный носитель, если на это нет письменного разрешения издательства "Диалектика".

Copyright © 2004 by Dialektika Computer Publishing.

All rights reserved including the right of reproduction in whole or in part in any form.

ISBN 5-8459-0692-X (рус.)

© Компьютерное изд-во "Диалектика", 2004

Оглавление

ЧАСТЬ I. ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ	19
Глава 1. Основные понятия теории вероятностей	20
Глава 2. Основные статистические методы	49
Глава 3. Анализ статистических зависимостей	78
ЧАСТЬ II. СРЕДСТВА EXCEL ДЛЯ СТАТИСТИЧЕСКОГО АНАЛИЗА	101
Глава 4. Статистические функции	102
Глава 5. Надстройка Пакет анализа	146
Глава 6. Дополнительные возможности Excel для проведения статистического анализа	193
Глава 7. Моделирование случайных величин	229
ЧАСТЬ III. АНАЛИЗ ОДНОМЕРНЫХ ВЫБОРОК	249
Глава 8. Предварительный анализ	250
Глава 9. Подбор распределения	286
Глава 10. Интервальное оценивание параметров распределения	307
Глава 11. Проверка гипотез о параметрах распределений	335
Глава 12. Сравнение одномерных выборок	349
ЧАСТЬ IV. СТАТИСТИЧЕСКИЙ АНАЛИЗ ЗАВИСИМОСТЕЙ	381
Глава 13. Корреляционный анализ	382
Глава 14. Сравнение зависимых выборок	400
Глава 15. Регрессионный анализ	417
Литература	427
Предметный указатель	429

Содержание

Предисловие	15
ЧАСТЬ I. ОСНОВНЫЕ ПОНЯТИЯ ТЕОРИИ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ	19
Глава 1. Основные понятия теории вероятностей	20
1.1. Понятия случайного события и случайной величины	20
1.1.1. Вероятности	21
1.1.2. Алгебра случайных событий	22
1.1.3. Условные вероятности	22
1.2. Распределения случайных величин	23
1.2.1. Дискретные случайные величины	23
1.2.2. Непрерывные случайные величины	25
1.2.3. Числовые характеристики случайных величин	25
1.2.4. Вероятностные неравенства	27
1.2.5. Двумерные распределения	28
1.3. Функции от случайных величин	29
1.3.1. Линейное преобразование случайных величин	30
1.3.2. Суммы случайных величин	30
1.3.3. Центральная предельная теорема	31
1.4. Примеры дискретных распределений	32
1.4.1. Равномерное дискретное распределение	32
1.4.2. Распределение Бернулли	32
1.4.3. Биномиальное распределение	33
1.4.4. Распределение Пуассона	34
1.4.5. Геометрическое распределение	34
1.4.6. Гипергеометрическое распределение	35
1.4.7. Отрицательное биномиальное распределение (распределение Паскаля)	35
1.5. Примеры непрерывных распределений	36
1.5.1. Равномерное непрерывное распределение	36
1.5.2. Треугольное распределение	37
1.5.3. Показательное (экспоненциальное) распределение	37
1.5.4. Нормальное распределение	38
1.5.5. Распределение "хи-квадрат"	39
1.5.6. Распределение Стюдента	40
1.5.7. F-распределение	41
1.5.8. Логарифмически нормальное распределение	42
1.5.9. Бета-распределение	43
1.5.10. Гамма-распределение	44

1.5.11. Распределение Вейбулла-Гнеденко	44
1.5.12. Распределения Пирсона	45
Глава 2. Основные статистические методы	49
2.1. Точечное оценивание параметров распределения	49
2.1.1. Несмещенность оценки	50
2.1.2. Эффективность оценки	51
2.1.3. Состоятельность оценки	51
2.2. Интервальное оценивание параметров распределения	52
2.3. Выборочные статистики и интервальные оценки	54
2.3.1. Статистика для оценивания математического ожидания	54
2.3.2. Статистика для оценивания дисперсии	56
2.3.3. Статистики для оценивания моментов	58
2.3.4. Статистики для оценивания коэффициентов асимметрии и эксцесса	58
2.3.5. Статистика для оценивания медианы	59
2.3.6. Оценки параметров нормального распределения	59
2.3.7. Оценка параметра p распределения Бернулли	61
2.3.8. Оценка параметра λ распределения Пуассона	63
2.3.9. Порядковые статистики	65
2.4. Проверка статистических гипотез	65
2.4.1. Критерии проверки гипотез о значениях параметров генеральной совокупности	68
2.4.2. Критерии сравнения значений параметров генеральных совокупностей	70
2.4.3. Критерии проверки гипотез о принадлежности распределения выборки классу распределений	75
Глава 3. Анализ статистических зависимостей	78
3.1. Общая модель статистических зависимостей	78
3.2. Задачи статистического анализа зависимостей	79
3.3. Корреляционный анализ	81
3.3.1. Анализ зависимостей между количественными переменными	81
3.3.2. Анализ зависимостей между порядковыми переменными	83
3.3.3. Анализ зависимостей между классификационными переменными	86
3.4. Регрессионный анализ	88
3.4.1. Выбор функции регрессии	88
3.4.2. Построение функции регрессии	90
3.4.3. Проверка адекватности функции регрессии	91
3.4.4. Статистические характеристики параметров функции регрессии	92
3.4.5. Прогнозирование	93
3.5. Дисперсионный анализ	94
3.5.1. Статистическая модель	94

3.5.2. Однофакторный дисперсионный анализ	95
3.5.3. Двухфакторный дисперсионный анализ	97

ЧАСТЬ II. СРЕДСТВА EXCEL ДЛЯ СТАТИСТИЧЕСКОГО АНАЛИЗА 101

Глава 4. Статистические функции	102
4.1. Функции для определения экстремальных значений выборки	102
4.1.1. Функции МАКС, МАКСА, МИН, МИНА	103
4.1.2. Функции НАИБОЛЬШИЙ и НАИМЕНЬШИЙ	103
4.2. Функции для работы с порядковыми статистиками	104
4.2.1. Функция КВАРТИЛЬ	104
4.2.2. Функция ПЕРСЕНТИЛЬ	105
4.2.3. Функция МЕДИАНА	106
4.2.4. Функция ПРОЦЕНТРАНГ	106
4.2.5. Функция РАНГ	107
4.3. Функции для вычисления средних	109
4.3.1. Функция СРГAM	109
4.3.2. Функция СРГЕОМ	109
4.3.3. Функции СРЗНАЧ и СРЗНАЧА	109
4.3.4. Функция УРЕЗСРЕДНЕЕ	110
4.4. Функции для вычисления геометрических характеристик распределения	110
4.4.1. Функция СКОС	110
4.4.2. Функция ЭКСЦЕСС	111
4.5. Функции для вычисления выборочной дисперсии и отклонения	111
4.5.1. Функции ДИСП и ДИСПА	112
4.5.2. Функции ДИСПР и ДИСПРА	112
4.5.3. Функция КВАДРОТКЛ	112
4.5.4. Функции СТАНДОТКЛОН и СТАНДОТКЛОНА	112
4.5.5. Функции СТАНДОТКЛОНП и СТАНДОТКЛОНПА	113
4.5.6. Функция СРОТКЛ	113
4.6. Функции для вычисления значений функций распределения	113
4.6.1. Функция ФРАСП	114
4.6.2. Функция БЕТАРАСП	114
4.6.3. Функция БИНОМРАСП	115
4.6.4. Функция ВЕЙБУЛЛ	115
4.6.5. Функция ГАММАРАСП	116
4.6.6. Функция ГИПЕРГЕОМЕТ	116
4.6.7. Функция ЛОГНОРМРАСП	117
4.6.8. Функция НОРМРАСП	117
4.6.9. Функция НОРМСТРАСП	117
4.6.10. Функция ОТРБИНОМРАСП	117
4.6.11. Функция ПУАССОН	118
4.6.12. Функция СТЫЮДРАСП	118
4.6.13. Функция ХИ2РАСП	119
4.6.14. Функция ЭКСПРАСП	119

4.7. Функции, обратные к функциям распределения	119
4.7.1. Функция ФРАОТОВР	120
4.7.2. Функция БЕТАОБР	121
4.7.3. Функция ГАММАОБР	121
4.7.4. Функция ЛОГНОРМОБР	121
4.7.5. Функция НОРМОБР	122
4.7.6. Функция НОРМСТОБР	122
4.7.7. Функция СТЬЮДРАСПОБР	122
4.7.8. Функция ХИ2ОБР	122
4.7.9. Функция КРИТБИНОМ	123
4.8. Функции для проверки статистических критериев	123
4.8.1. Функция ZТЕСТ	124
4.8.2. Функция ТТЕСТ	124
4.8.3. Функция ФТЕСТ	126
4.8.4. Функция ХИ2ТЕСТ	127
4.9. Функции для построения уравнения регрессии и прогнозирования	128
4.9.1. Функция ЛИНЕЙН	129
4.9.2. Функции НАКЛОН и ОТРЕЗОК	131
4.9.3. Функция СТОПНУХ	132
4.9.4. Функция ПРЕДСКАЗ	133
4.9.5. Функция ТЕНДЕНЦИЯ	133
4.9.6. Функция ЛГРФПРИБЛ	134
4.9.7. Функция РОСТ	135
4.10. Функции для вычисления ковариации и коэффициента корреляции	136
4.10.1. Функция КОВАР	136
4.10.2. Функция КОРРЕЛ	137
4.10.3. Функция ПИРСОН	137
4.10.4. Функция КВПИРСОН	138
4.10.5. Функции ФИШЕР и ФИШЕРОБР	139
4.11. Дополнительные функции	139
4.11.1. Функция ВЕРОЯТНОСТЬ	140
4.11.2. Функция ДОВЕРИТ	140
4.11.3. Функция МОДА	141
4.11.4. Функция ЧАСТОТА	141
4.12. Вспомогательные функции	142
4.12.1. Функция ГАММАНЛОГ	142
4.12.2. Функция НОРМАЛИЗАЦИЯ	142
4.12.3. Функция ПЕРЕСТ .	143
4.12.4. Функции СЧЁТ и СЧЁТЗ	143
4.13. Функции для генерирования равномерно распределенных случайных чисел	143
4.13.1. Функция С Л ЧИС	144
4.13.2. Функция СЛУЧМЕЖДУ	144

Глава 5. Надстройка Пакет анализа	146
5.1. Описательная статистика	149
5.1.1. Опции диалогового окна Описательная статистика	151
5.2. Гистограмма	151
5.2.1. Опции диалогового окна Гистограмма	152
5.3. Генерация случайных чисел	154
5.3.1. Опции диалогового окна Генерация случайных чисел	155
5.4. Выборка	160
5.4.1. Опции диалогового окна Выборка	160
5.5. Ранг и персентиль	161
5.6. Двухвыборочный z-тест для средних	161
5.7. Двухвыборочный t-тест с одинаковыми дисперсиями	165
5.8. Двухвыборочный t-тест с различными дисперсиями	167
5.9. Парный двухвыборочный t-тест для средних	169
5.10. Двухвыборочный F-тест для дисперсий	172
5.11. Однофакторный дисперсионный анализ	173
5.12. Двухфакторный дисперсионный анализ с повторениями	175
5.13. Двухфакторный дисперсионный анализ без повторений	177
5.14. Корреляция	179
5.15. Ковариация	180
5.16. Регрессия	181
5.17. Скользящее среднее	187
5.18. Экспоненциальное сглаживание	188
5.19. Анализ Фурье	189
Глава 6. Дополнительные возможности Excel для проведения статистического анализа	193
6.1. Массивы и формулы массивов	193
6.1.1. Редактирование формул массивов	196
6.1.2. Массивы констант	196
6.1.3. Поименованные массивы и диапазоны	197
6.1.4. Примеры использования формул массивов	200
6.1.5. Матричные вычисления	203
6.1.6. Функции суммирования	204
6.2. Диаграммы	206
6.2.1. Линии тренда	207
6.2.2. Планки погрешностей	210
6.2.3. Построение гистограмм и функций распределения	
*	
дискретных случайных величин	212
6.2.4. Гистограммы с перекрытием	215
6.3. Надстройка Поиск решения	217
6.3.1. Задачи оптимизации и средство Поиск решения	218
6.3.2. Задачи, решаемые средством Поиск решения	224
6.3.3. Примеры применения средства Поиск решения	225

Глава 7. Моделирование случайных величин	229
7.1. Средства Excel для генерирования случайных чисел	229
7.2. Метод обратных функций моделирования случайных величин	234
7.3. Метод суперпозиций	238
7.4. Метод отбора	242
7.5. Моделирование многомерных случайных величин	244
7.5.1. Моделирование зависимых случайных величин с известным коэффициентом корреляции	245
ЧАСТЬ III. АНАЛИЗ ОДНОМЕРНЫХ ВЫБОРОК	249
Глава 8. Предварительный анализ	250
8.1. Цензурирование	250
8.1.1. Цензурирования на основе построения доверительных интервалов	251
8.1.2. Непараметрическое цензурирование	257
8.1.3. Винзоризация выборки	258
8.2. Преобразование данных	263
8.2.1. Преобразование квадратного корня	263
8.2.2. Логарифмическое преобразование	265
8.2.3. Стандартизирующее преобразование	267
8.3. Построение гистограмм, полигонов и эмпирических функций распределения	267
8.3.1. Построение гистограммы и эмпирической функции распределения для дискретных случайных величин	268
8.3.2. Построение гистограммы и полигона для непрерывных распределений	273
8.4. Вычисление точечных оценок параметров распределения	278
8.4.1. Точечные оценки дискретного распределения	283
8.4.2. Вычисление моды для непрерывных распределений	285
Глава 9. Подбор распределения	286
9.1. Предварительное определение класса распределения	286
9.1.1. Построение пробит-графиков	288
9.2. Подбор функции распределения на основе числовых характеристик выборки	291
9.2.1. Критерии отклонения распределения от нормального	293
9.2.2. Критерий отклонения от распределения Пуассона	296
9.3. Критерий χ^2	297
9.3.1. Критерий χ^2 для дискретных распределений	297
9.3.2. Критерий $\%^2$ для непрерывных распределений	299
9.4. Критерий Колмогорова	304
Глава 10. Интервальное оценивание параметров распределения	307
10.1. Общие доверительные интервалы для математического ожидания	307
10.1.1. Общая модель при известной дисперсии	307

10.1.2. Одномодальное симметричное распределение при известной дисперсии	308
10.1.3. Общая модель с неизвестной дисперсией	308
10.2. Общий доверительный интервал для дисперсии	310
10.3. Интервальные оценки параметров нормального распределения	312
10.3.1. Интервальные оценки для неизвестного математического ожидания при известной дисперсии	312
10.3.2. Интервальные оценки для неизвестного математического ожидания при неизвестной дисперсии	313
10.3.3. Интервальные оценки для неизвестной дисперсии при известном математическом ожидании	315
10.3.4. Интервальные оценки для неизвестной дисперсии при неизвестном математическом ожидании	315
10.4. Оценка параметров логарифмически нормального распределения	317
10.5. Оценка параметра показательного распределения	318
10.6. Оценка параметров гамма-распределения	319
10.6.1. Оценка параметра A при известном параметре a	320
10.6.2. Оценка параметра a при известном параметре X	321
10.6.3. Совместная оценка параметров α и A	322
10.7. Оценка параметров равномерного распределения	323
10.7.1. Оценка границы равномерного распределения	323
10.7.2. Оценка обеих границ равномерного распределения	324
10.8. Оценки параметра распределения Бернулли	324
10.8.1. Оценивание вероятности p по одному эксперименту	325
10.8.2. Оценивание вероятности p по нескольким экспериментам	327
10.8.3. Применение преобразования арксинуса	328
10.9. Оценка параметра распределения Пуассона	329
10.10. Оценки параметра геометрического распределения	331
10.11. Доверительные интервалы для квантилей	333
Глава 11. Проверка гипотез о параметрах распределений	335
11.1. Критерии проверки гипотез о параметрах нормального распределения	335
11.1.1. Критерий проверки значения математического ожидания нормальной совокупности	335
11.1.2. Критерий проверки значения дисперсии нормальной совокупности	337
11.2. Проверка гипотезы о значении параметра показательного распределения	339
11.3. Проверка гипотезы о значении параметра биномиального распределения	341
11.3.1. Использование биномиального распределения	341
11.3.2. Асимптотический критерий	343
11.4. Критерии проверки гипотез о значении медианы	343
11.4.1. Критерий знаков	344
11.4.2. Критерий знаковых рангов Уилкоксона	346

Глава 12. Сравнение одномерных выборок	349
12.1. Сравнение выборочных распределений	349
12.1.1. Непараметрический критерий медианы	350
12.1.2. Критерий Уилкоксона-Манна-Уитни	355
12.1.3. Критерий Краскала-Уоллиса	357
12.1.4. Критерий серий Вальда-Вольфовица	359
12.1.5. Критерий χ^2	360
12.1.6. Критерий Смирнова	362
12.2. Доверительные интервалы для параметров распределений	364
12.2.1. Доверительный интервал для разности средних нормальных совокупностей (равные дисперсии)	364
12.2.2. Доверительный интервал для разности средних нормальных совокупностей (разные дисперсии)	365
12.2.3. Доверительный интервал для отношения дисперсий нормальных совокупностей	366
12.2.4. Доверительный интервал для разности двух биномиальных вероятностей	367
12.3. Проверка гипотез о параметрах распределений	368
12.3.1. Проверка гипотез о математических ожиданиях нормальных распределений	368
12.3.2. Проверка гипотез о дисперсиях нормальных распределений	374
12.3.3. Непараметрический критерий Ансари-Бредли проверки гипотезы о равенстве дисперсий	378
12.3.4. Проверка гипотез о равенстве биномиальных вероятностей	380
 ЧАСТЬ IV. СТАТИСТИЧЕСКИЙ АНАЛИЗ ЗАВИСИМОСТЕЙ	 381
Глава 13. Корреляционный анализ	382
13.1. Критерии независимости	382
13.1.1. Критерий независимости на основе преобразования Фишера	383
13.1.2. Критерий независимости для двумерных нормальных совокупностей	384
13.1.3. Критерий независимости на основе рангового коэффициента корреляции Спирмена	385
13.1.4. Критерий независимости на основе рангового коэффициента корреляции Кендалла	386
13.1.5. Критерий независимости для многомерных выборок	389
13.1.6. Критерий независимости на основе таблиц сопряженности	390
13.2. Оценивание коэффициента корреляции	393
13.2.1. Доверительные интервалы для коэффициента корреляции	393
13.2.2. Доверительные интервалы для коэффициента корреляции нормальной совокупности	394
13.3. Критерии проверки гипотез о значениях коэффициента'	396
корреляции	396
13.3.1. Критерий проверки значения коэффициента корреляции	396

13.3.2. Критерий проверки равенства двух коэффициентов корреляции	397
13.3.3. Критерий проверки равенства нескольких коэффициентов корреляции	399
Глава 14. Сравнение зависимых выборок	400
14.1. Доверительные интервалы для разности математических ожиданий нормальных совокупностей	400
14.1.1. Доверительный интервал для разности математических ожиданий	400
14.1.2. Доверительный интервал для математических ожиданий нескольких совокупностей	401
14.2. Критерии проверки гипотез о равенстве математических ожиданий	403
14.2.1. Парный критерий Стьюдента	404
14.2.2. Непараметрический критерий знаков	405
14.2.3. Непараметрический критерий Уилкоксона	407
14.3. Дисперсионный анализ для зависимых выборок	408
14.3.1. Двухфакторный дисперсионный анализ	409
14.3.2. Двухфакторный дисперсионный анализ Фридмана	411
14.3.3. Критерий множественных сравнений Шеффе для зависимых выборок	415
Глава 15. Регрессионный анализ	417
15.1. Построение функции регрессии	418
15.2. Адекватность уравнения регрессии	420
15.3. Доверительные интервалы и проверка гипотез для коэффициентов функции регрессии	422
15.4. Доверительный интервал для значения прогноза	423
Литература	427
Предметный указатель	429

Предисловие

Сегодня в различных сферах общественной жизни к статистическим методам проявляется повышенный интерес как к одному из важнейших аналитических инструментов для поддержки процессов принятия решений. Статистикой пользуются все: от бизнесменов, стремящихся оптимизировать прибыль от инвестиций, до политиков, желающих предсказать исход выборов, или социологов, оценивающих доверие избирателей к этим политикам, не говоря уже о традиционных областях применения математической статистики — науке, технике, экономике. Очевидно, что, как правило, статистическими методами в своей деятельности пользуются не профессионалы-статистики (где набрать столько профессионалов!), а "обычные" профессионалы в своей области, которые, возможно, когда-то "проходили" в своих университетах курс математической статистики, но "это было так давно, что стало неправдой".

Мой достаточно большой опыт применения статистических методов в совместной работе с биологами, медиками и в последние годы с экономистами показывает, что распространенное мнение о статистике как об одной из разновидностей лжи идет от неправомерного применения тех или иных статистических методов в конкретных ситуациях. Даже общеупотребительный и "безопасный" критерий Стьюдента, примененный бездумно, например, к выборкам из дискретных генеральных совокупностей, может в некоторых случаях показать удивительные результаты. С другой стороны, почти во всей литературе по математической статистике, включая практические руководства, материал излагается таким образом, что сначала идет "теория", например основы метода максимального правдоподобия, а затем в качестве иллюстрации к "теории" предлагается несколько практических методов. В таком случае практику весьма сложно выбрать необходимые методы проведения статистического анализа, сравнить эти методы и тем более обосновать их применение. (Редким исключением в общем ряду такой статистической литературы является книга Дж. Полларда *Справочник по вычислительным методам статистики*, в которой представлены практические методы статистики и описание области применимости каждого из них.)

Эта книга задумывалась и написана как "сборник рецептов" статистических методов, которые часто используются на практике и сравнительно просто реализуются в электронной таблице Excel. Для каждого приведенного метода *четко описана статистическая модель*, в рамках которой его можно применять. Кроме того, методы сгруппированы по типу исходных данных, предъявляемых для статистического анализа. Таким образом, отдельно описаны методы для анализа одномерных выборок, отдельно — для зависимых наблюдений и т.д. Методы представлены в таком виде, чтобы их могли легко отобрать для своих потребностей и сравнительно просто реализовать практики (необязательно профессионалы-статистики), которым необходимо самостоятельно провести статистический анализ своих данных.

В этой связи необходимо отметить выбор электронной таблицы Excel как средства реализации методов статистического анализа. Существует множество специализированных программных средств для статистических расчетов: отечественные STADIA, СИГАМД, ОЛИМП:СтатЭксперт или зарубежные STATGRAPHICS,

STATISTICA, SPSS и общематематические пакеты (например, Mathcad, Mathlab, Maple), которые также имеют встроенные статистические средства. Но наибольшее распространение как средство проведения различных расчетов, в том числе и статистических, в настоящее время получили электронные таблицы, среди которых безусловным лидером является Microsoft Excel. Эта электронная таблица входит в пакет Microsoft Office, который установлен практически на каждом компьютере. Microsoft Excel имеет достаточное количество встроенных статистических средств, включая надстройку Пакет анализа и порядка 80 статистических функций. Это обусловило выбор Excel в качестве основного средства для проведения статистического анализа. Несмотря на то что в книге все примеры реализованы в Excel 2002, они без существенных изменений могут быть перенесены на другие версии Excel, начиная с Excel 97 и заканчивая Excel 2003.

Хотя книга задумывалась только как сборник статистических методов, оказалось невозможным обойтись без вводной части, посвященной основам теории вероятностей и математической статистики, и специальной части, описывающей статистические возможности Excel. Поэтому книга состоит из четырех частей. В части I, *Основные понятия теории вероятностей и математической статистики*, приводятся основные понятия и сведения из теории вероятностей и математической статистики. Весь материал этой части представлен конспективно; здесь приведены все необходимые базовые понятия, определения, теоремы и статистические модели, которые позволят читателю вполне осознанно и продуктивно применять статистические методы, описанные в последующих частях книги. Конечно, эта часть совсем не предназначена для того, чтобы по ней изучать такую обширную и насыщенную область математики (хотя некоторые темы освещены достаточно подробно), как теория вероятностей и математическая статистика. Ее можно использовать как справочное пособие, к которому рано или поздно будет вынужден обратиться как практик-"нестатистик", который использует статистический анализ в своей работе, так и специалист-статистик (у любого специалиста рано или поздно возникает необходимость вернуться к "истокам" — базовым понятиям). Кроме того, материал этой части используется в части II для ссылок при описании статистических средств Excel.

В части II, *Средства Excel для статистического анализа*, описаны возможности Excel для проведения статистического анализа. Предполагается, что читатель знаком с основами работы в этой электронной таблице хотя бы в следующем объеме: он может вводить и редактировать данные, создавать формулы, использовать функции, строить диаграммы и графики, форматировать рабочий лист и т.п. Это базовые навыки работы с Excel, которые известны каждому, кто прослушал курс информатики и вычислительной техники (и при этом, конечно, усвоил необходимые знания) в любом вузе любого профиля. В этой части достаточно полно описаны статистические функции и средства, предоставляемые надстройкой Пакет анализа. К сожалению, справочная система Excel настолько неполно и невнятно (и даже с ошибками!) представляет эти функции и средства, что необходимость их полного описания очевидна. (Следует отметить, что в Excel 2003 справочная система написана более профессионально, при этом исправлены некоторые ошибки.) Кроме статистических функций и средств пакета анализа, в данной части рассмотрены общие средства и надстройки Excel, которые "не заявлены" как имеющие непосредственное отношение к статистическим методам, но которые также можно использовать в статистическом анализе. Это формулы массивов, специального вида

диаграммы и графики, а также надстройка Поиск решения. В конце части описаны способы моделирования случайных величин в Excel.

В части III, *Анализ одномерных выборок*, показана практическая реализация методов статистического анализа одномерных независимых выборок, рассмотрены вопросы предварительной обработки данных и подбора распределений по имеющимся выборочным значениям, а также приведены методы интервального оценивания параметров распределений и критерии проверки гипотез о значениях этих параметров. Последняя глава части посвящена сравнению распределений нескольких одномерных выборок.

В части IV, *Статистический анализ зависимостей*, описаны методы анализа статистических зависимостей, которые включают в себя широкий спектр статистических алгоритмов. Здесь рассмотрены методы корреляционного анализа, способы построения доверительных интервалов и критерии проверки гипотез о значениях коэффициента корреляции, а также показаны методы сравнения параметров распределений зависимых компонентов многомерных выборок. В последней главе описан ряд задач, связанных с построением регрессий, начиная с общей вычислительной схемы определения коэффициентов уравнений регрессии и заканчивая критериями проверки адекватности построенного уравнения регрессии. Хотя число рассмотренных в этой части методов достаточно велико и сами методы весьма громоздки, часть получилась на удивление небольшой. "Виной" этому Excel, в которой есть практически все средства, необходимые для реализации данных методов.

В конце книги приведен небольшой список литературы, на которую есть ссылки в тексте или которая может дополнить определенные темы, освещенные недостаточно полно.

Я буду признателен всем, кто поделится своими соображениями по улучшению содержания книги и стиля изложения материала, а также укажет на возможные ошибки (к сожалению, в книгах, содержащих более ста формул, вероятность ошибок всегда отлична от нуля). Мой адрес электронной почты — aminko@dialektika.com.

А.Л. Минько

От издательства "Диалектика"

Вы, читатель этой книги, и есть главный ее критик. Мы ценим ваше мнение и хотим знать, что было сделано нами правильно, что можно было сделать лучше и что еще вы хотели бы увидеть изданным нами. Нам интересно услышать и любые другие замечания, которые вам хотелось бы высказать в наш адрес.

Мы ждем ваших комментариев и надеемся на них. Вы можете прислать нам бумажное или электронное письмо либо просто посетить наш Web-сервер и оставить свои замечания там. Одним словом, любым удобным для вас способом дайте нам знать, нравится ли вам эта книга, а также выскажите свое мнение о том, как сделать наши книги более интересными для вас.

Посылая письмо или сообщение, не забудьте указать название книги и ее авторов, а также ваш обратный адрес. Мы внимательно ознакомимся с вашим мнением и обязательно учтем его при отборе и подготовке к изданию последующих книг. Наши координаты:

E-mail: info@dialektika.com

WWW: <http://www.dialektika.com>

Информация для писем из:

России: 115419, Москва, а/я 783

Украины: 03150, Киев, а/я 152 -

Основные понятия теории вероятностей и математической статистики

В этой части...

Глава 1. Основные понятия теории вероятностей

Глава 2. Основные статистические методы

Глава 3. Анализ статистических зависимостей

В главе 1 этой части приводятся основные понятия и сведения из теории вероятностей. Материал по математической статистике представлен в двух главах: в главе 2 приводятся общие сведения по статистике (большая часть этой главы посвящена интервальному оцениванию и проверке гипотез), в главе 3 описывается статистический анализ зависимостей. Весь материал представлен конспективно и предназначен скорее для того, чтобы "освежить" в памяти читателя теорию вероятностей и математическую статистику, но, конечно, совсем не для того, чтобы по этим главам изучать такую обширную и насыщенную область математики (хотя некоторые темы освещены достаточно подробно). Вместе с тем любой практик-"нестатистик", который использует статистический анализ в своей работе, найдет здесь все необходимые базовые понятия, определения, теоремы и статистические модели, которые дадут ему возможность вполне осознанно и продуктивно применить статистические методы, описанные в последующих частях книги. Специалист-статистик может использовать материал этой части в качестве справочного пособия.

Основные понятия теории вероятностей

В данной главе приводятся основные понятия и сведения из теории вероятностей, необходимые для изложения основ математической статистики в последующих главах. Значительная часть главы посвящена примерам вероятностных распределений, которые часто встречаются при проведении статистического анализа, в том числе приведено полное описание системы распределений Пирсона. Эту часть главы можно использовать как справочный материал по вероятностным распределениям.

1.1. Понятия случайного события и случайной величины

Среди основных понятий теории вероятностей и математической статистики понятия *опыт* (эксперимент) и *событие* являются фундаментальными. Будем называть опытом наблюдение какого-либо явления при выполнении некоторого комплекса условий, который должен каждый раз строго выполняться при повторении данного опыта. Наблюдение того же явления при другом комплексе условий будет уже другим опытом. Результат *случайного опыта* не известен до его окончания. Далее будем иметь дело только со случайным опытом.

Результаты случайного опыта можно охарактеризовать качественно и количественно. Качественная характеристика опыта состоит в регистрации какого-либо факта. Любой такой факт называется *случайным событием*. При этом говорят, что "событие произошло (появилось)" или "событие не произошло (не появилось)" в результате случайного опыта.

Примерами событий могут служить выпадение решки при бросании монеты или цифры "3" при бросании игральной кости, отказ прибора в заданном интервале времени, попадание или промах при выстреле, получение m попаданий при n выстрелах и т.д. Итак, случайным событием (или просто "событием") называется всякий факт, который в результате опыта может произойти или не произойти.

Количественная характеристика опыта состоит в определении значений некоторых величин, полученных в результате опыта. Величины, которые могут принимать в результате опыта различные значения, причем до опыта невозможно предвидеть, какими именно они будут, называются *случайными величинами*. Примерами случайных величин могут служить как результаты, так и ошибки измерений, время безотказной работы прибора или системы, рост и вес наугад выбранного человека, число попаданий при n выстрелах и т.д.

С каждой случайной величиной можно связать различные случайные события. Типичным событием, связанным со случайной величиной, является событие, состоящее в том, что эта случайная величина примет в результате опыта какое-либо значение, принадлежащее заданному числовому множеству. Кратко такое событие называется попаданием случайной величины в данное множество значений.

1.1.1. Вероятности

Естественно сравнивать события по тому, как часто каждое из них появляется при повторении данного опыта. Если при повторении опыта одно событие появляется чаще, чем другое, то говорят, что первое событие *вероятнее* второго. При этом ясно, что для сравнения событий необходимо предположить, что данный опыт можно проводить сколько угодно раз при соблюдении одного и того же комплекса условий.

Частотой появления события A называется отношение числа его появлений к числу всех проведенных опытов. Таким образом, если в n опытах событие A появилось m раз, то частота его появления в данной серии опытов равна m/n .

Важным экспериментально установленным фактом является свойство устойчивости частот. При увеличении числа опытов частоты событий колеблются около некоторых чисел, не зависящих ни от количества, ни от серии опытов, причем частоты неограниченно приближаются к этим числам, когда число опытов стремится к бесконечности. (В теории вероятностей этот факт называется *законом больших чисел*. В качестве иллюстрации на рис. 1.1 показана рабочая книга Excel, где смоделировано 1 000 подбрасываний монеты и построен график частот выпадения герба.) Эти числа естественно связать с каждым событием, происходящим в случайном опыте. Они называются *вероятностями* и в теории вероятностей определяются чисто аксиоматически. Вероятность события A обозначается как $P(A)$ и может принимать любые значения от нуля до единицы: $0 < P(A) < 1$.

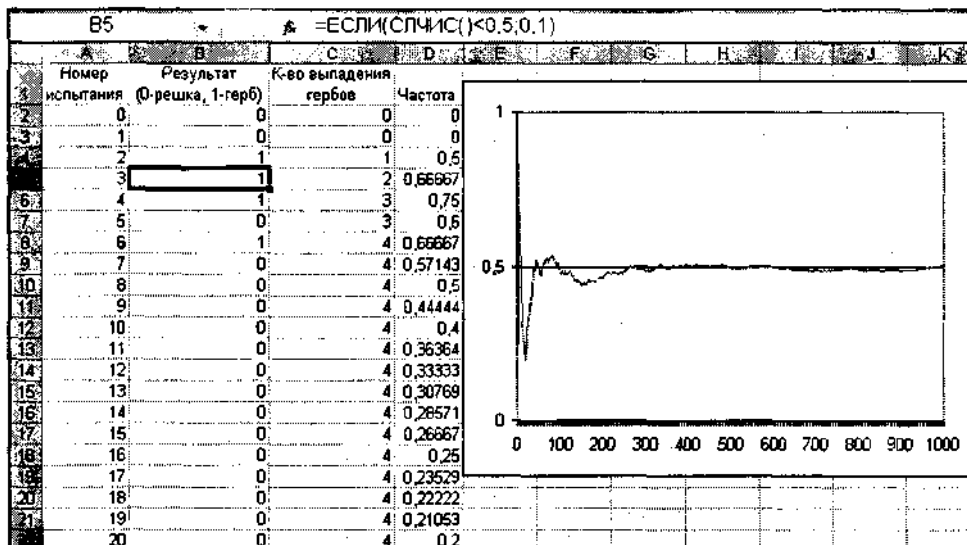


Рис. 1.1. Модель подбрасывания монеты и график частоты выпадения герба

1.1.2. Алгебра случайных событий

Приведем некоторые важные определения и теоремы алгебры случайных событий.

Различают следующие виды случайных событий. *Достоверным* называется событие U , которое в результате опыта непременно должно произойти, в этом случае $P(U) = 1$. *Невозможным* называется событие V , которое в результате опыта не может произойти никогда; тогда $P(V) = 0$. Событие \bar{A} называется *противоположным* событию A , если оно состоит в непоявлении события A . Сумма вероятностей противоположных событий всегда равна единице: $P(\bar{A}) + P(A) = 1$. Например, при подбрасывании монеты может произойти только одно из двух событий (выпадение орла или выпадение решки), которые не могут произойти одновременно. Поэтому данные события противоположны.

Несколько событий в данном опыте называются *несовместными* или *взаимоисключающими*, если никакие два из них не могут появиться вместе. Классический пример несовместных событий: 6 событий, состоящих в том, что при бросании игрального кубика появятся цифры 1, 2, 3, 4, 5 или 6 соответственно. Сумма вероятностей всех несовместных событий, связанных с тем или иным опытом, равна единице (в таком случае говорят, что эти события составляют *полную группу* событий). Если A_1, A_2, \dots, A_n — несовместные события, то

$$P(A_1 \text{ или } A_2 \text{ или } \dots \text{ или } A_n) = P(A_1) + P(A_2) + \dots + P(A_n).$$

Это равенство называется *теоремой сложения вероятностей*.

В левой части последнего равенства записана вероятность *суммы событий*: суммой (объединением) двух событий A и B называется событие " A или B " (также обозначается как $A + B$ или $A \cup B$), происходящее тогда и только тогда, когда происходит или событие A , или событие B . Аналогично определяется сумма любого числа событий. В алгебре случайных событий вводится еще одна операция над событиями. *Произведением* (пересечением) событий A и B называется событие " A и B " (также обозначается как $A \cap B$ или $A \wedge B$), происходящее тогда и только тогда, когда происходит и событие A , и событие B . Подобным образом определяется произведение любого числа событий.

События A и B называются *независимыми*, если появление одного из них не меняет вероятности появления другого. Например, независимыми будут события "при первом бросании игрального кубика откроется цифра 2" и "при втором бросании игрального кубика откроется цифра 5". Если A_1, A_2, \dots, A_n — взаимно независимые случайные события, то

$$P(A_1 \text{ и } A_2 \text{ и } \dots \text{ и } A_n) = P(A_1) \times P(A_2) \times \dots \times P(A_n).$$

Это равенство называется *теоремой умножения вероятностей*.

1.1.3. Условные вероятности

Условную вероятность события A при условии, что произошло событие B , обозначают как $P(A|B)$. Приведем формулу, связывающую вероятность совместного появления событий A и B и условные вероятности этих событий:

$$P(A \text{ и } B) = P(A) P(B|A) = P(B) P(A|B).$$

Если события A и B независимы, то $P(A|B) = P(A)$ и $P(B|A) = P(B)$.

Пример 1. Игрок бросает пять раз симметричную монету. Считая, что подбрасывания независимы, какова вероятность события, что герб выпадет точно два раза?

Герб выпадет точно два раза только при следующих десяти возможных последовательностях выпадения герба или решки (Г — герб, Р — решка):

ГГРРР, ГРГРР, ГРРГР, ГРРРГ, РГГРР, РГРГР, РГРРГ, РРГГР, РРГРГ, РРРГГ.

При любом подбрасывании монеты вероятность выпадения герба и вероятность выпадения решки равны $1/32$. Пять подбрасываний монеты независимы. Из теоремы умножения вероятностей следует, что вероятность получения последовательности ГГРРР можно подсчитать следующим образом:

$$P(\text{ГГРРР}) = P(\text{Г}) \cdot P(\text{Г}) \cdot P(\text{Р}) \cdot P(\text{Р}) \cdot P(\text{Р}) = (1/2)^5 = 1/32.$$

Для каждой из остальных десяти последовательностей вероятность также равна $1/32$. Каждая из этих десяти последовательностей содержит по два выпадения герба, и эти последовательности являются взаимоисключающими. Из теоремы сложения вероятностей следует, что искомая вероятность представляет собой сумму этих десяти равных между собой вероятностей, т.е. она равна $10/32$.

Аналогично можно показать, что вероятность события "герб не выпадет ни разу" равна $1/32$; вероятность того, что герб выпадет ровно один раз, равна $5/32$; вероятность выпадения герба ровно три раза равна $10/32$; вероятности того, что герб выпадет точно 4 и 5 раз, равны $5/32$ и $1/32$ соответственно. Все шесть перечисленных выше событий являются взаимоисключающими и образуют полную группу событий. Как не трудно подсчитать, сумма их вероятностей равна 1.

1.2. Распределения случайных величин

Если случайное событие полностью характеризуется вероятностью появления этого события, а совокупность случайных событий можно описать с помощью вероятностей этих событий и теорем алгебры случайных событий, то описание вероятностных свойств случайных величин является более сложной задачей. Напомним, что случайной называется величина, которая в результате опыта может принимать то или иное значение (заранее неизвестно, какое именно).

Вероятностные свойства случайных величин описываются *законом распределения*, т.е. соотношением, устанавливающим связь между возможными значениями случайной величины и соответствующими им вероятностями. Закон распределения может иметь различные формы.

Различают дискретные и непрерывные случайные величины.

1.2.1. Дискретные случайные величины

Дискретной случайной величиной называют величину, принимающую только конечное или счетное множество значений.

Для описания дискретной случайной величины X , принимающей конечное множество значений, часто применяется таблица вида

x_i	x_1	x_2	...	x_{n-1}	x_n
$P(X = x_i)$	p_1	p_2	...	p_{n-1}	p_n

Здесь X_i — возможные значения случайной величины X , $p_i = P(X = x_i)$ — вероятность события, что случайная величина X примет значение x_i ($1 \leq i \leq n$). Отметим, что

$$2>_{\cdot} = i, / \backslash (x < u) =] >_{\cdot},$$

В последнем выражении суммирование ведется по всем таким номерам i , что $x_i < u$. Совокупность вероятностей $p_i = P(X = x_i)$ часто называют *функцией вероятностей*, а вероятность $P(X < u)$ обозначают как $F(u)$ и называют *функцией распределения* случайной величины X . Она является неубывающей разрывной ступенчатой функцией, принимающей значения в интервале от 0 до 1.

Пример 2. Как и в примере 1, игрок пять раз подбрасывает симметричную монету. Обозначим через X случайную величину, равную числу выпадения герба в серии подбрасывания монеты. Случайная величина X может принимать значения 0, 1, 2, 3, 4 и 5. Вероятность того, что случайная величина X примет какое-либо из этих значений, определена в примере 1. Составим таблицу распределения этой случайной величины. (На рис. 1.2 показано графическое представление этого распределения.)

x_i	0	1	2	3	4	5
$P(X = x_i)$	1/32	5/32	10/32	10/32	5/32	1/32

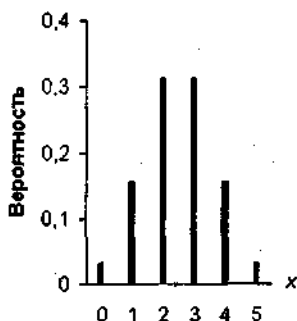


Рис. 1.2. Распределение вероятностей

Приведем значения функции распределения случайной величины X :

$$F(u) = P(X < u) = \begin{cases} 0, & u < 0, \\ p_1 = 1/32, & 0 \leq u < 1, \\ p_1 + p_2 = 6/32, & 1 \leq u < 2, \\ p_1 + p_2 + p_3 = 16/32, & 2 \leq u < 3, \\ p_1 + p_2 + p_3 + p_4 = 26/32, & 3 \leq u < 4, \\ p_1 + p_2 + p_3 + p_4 + p_5 = 31/32, & 4 \leq u < 5, \\ 1, & u \geq 5. \end{cases}$$

Графиком функции $F(u)$ будет возрастающая ступенчатая функция со скачками в точках $x = 1, 2, 3, 4, 5$, показанная на рис. 1.3.

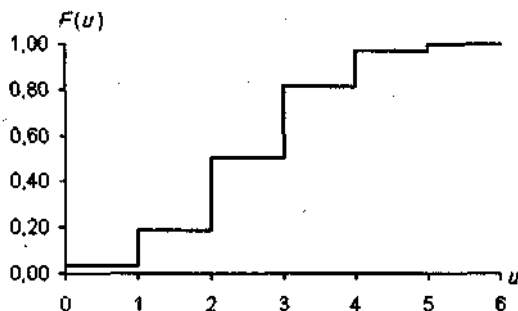


Рис. 1.3. График функции распределения

Примеры других дискретных распределений приведены в разделе 1.4.

1.2.2. Непрерывные случайные величины

Непрерывной случайной величиной называется случайная величина, возможные значения которой непрерывно заполняют какой-либо интервал (возможно, бесконечный). Для непрерывной случайной величины X также в качестве закона распределения выступает функция распределения $F(u)$, численно равная вероятности того, что случайная величина X окажется меньше заданного числа u , т.е. $F(u) = P(X < u)$. Функция $F(u)$ — непрерывная функция, неубывающая и принимающая значения в интервале от 0 до 1, причем $F(-\infty) = 0$ и $F(+\infty) = 1$.

Отметим, что распределение непрерывной случайной величины невозможно задать с помощью вероятностей отдельных значений подобно распределениям дискретных случайных величин, поскольку $P(X = x) = 0$ для любого значения x . Но если функция $F(u)$ дифференцируемая, то можно определить вероятность попадания случайной величины X в какой-либо малый интервал длиной dx , примаыкающий к точке x , и при этом $P(x \leq X < x + dx) = f(x)dx$, где $f(x)$ — производная функции $F(u)$ в точке x . Функция $f(x)$ называется *плотностью вероятности* случайной величины X . Она может принимать только неотрицательные значения. Из определения плотности вероятности следует, что

$$F(u) = \int_{-\infty}^u f(x)dx, \quad \int_{-\infty}^{+\infty} f(x)dx = 1, \quad P(a \leq X < b) = \int_a^b f(x)dx = F(b) - F(a).$$

Если случайная величина X может принимать только, например, положительные значения, то для такой случайной величины значения обеих функций $F(x)$ и $f(x)$ при отрицательных x должны быть нулевыми.

Примеры непрерывных случайных величин приведены в разделе 1.5.

1.2.3. Числовые характеристики случайных величин

Закон распределения полностью характеризует случайную величину. Чтобы определить закон распределения случайной величины, достаточно задать ее плотность вероятности или функцию распределения. Однако такая полная, исчерпывающая характеристика случайной величины довольно сложна. Между тем для решения многих задач практически вовсе не нужно знать распределение случайной величины, а достаточно знать лишь некоторые числа, характеризующие рас-

пределение, так называемые *числовые характеристики* случайной величины. Например, для грубого описания распределения случайной величины можно ограничиться ее средним значением и величиной разброса возможных значений.

Из числовых характеристик наиболее часто используются *моменты* случайной величины. Первый момент называется *математическим ожиданием* (или *средним* случайной величины) и вычисляется по одной из следующих формул (первая формула применяется для дискретных случайных величин, а вторая — для непрерывных):

$$MX = \sum_i x_i p_i, \quad MX = \int_{-\infty}^{+\infty} xf(x)dx.$$

Величина MX характеризует среднее положение значений случайной величины X .

Второй *центральный момент* (т.е. момент относительно математического ожидания MX) характеризует разброс значений случайной величины вокруг значения MX и называется *дисперсией*. Дисперсия DX (часто также используют обозначение σ^2 или σ_x^2) вычисляется по формулам (первая формула применяется для дискретных случайных величин, а вторая — для непрерывных)

$$DX = M(X - MX)^2 = \sum_i (x_i - MX)^2 p_i = \sum_i x_i^2 p_i - (MX)^2,$$

$$DX = M(X - MX)^2 = \int_{-\infty}^{+\infty} (x - MX)^2 f(x)dx = \int_{-\infty}^{+\infty} x^2 f(x)dx - (MX)^2.$$

На практике иногда используют моменты более высокого порядка, но, как правило, не выше четвертого. *Центральный момент r -го порядка μ_r* определяется как математическое ожидание от случайной величины $(X - MX)^r$ и вычисляется по формулам

$$\mu_r = M(X - MX)^r = \sum_i (x_i - MX)^r p_i,$$

$$\mu_r = M(X - MX)^r = \int_{-\infty}^{+\infty} (x - MX)^r dx,$$

соответствующим дискретному и непрерывному случаям. В этих обозначениях $DX = \mu_2$.

Для симметричных распределений все центральные моменты нечетного порядка равны нулю. Они положительны, если распределение асимметрично и имеет длинный "хвост" справа от математического ожидания (примером такого распределения может служить JP -распределение, описанное ниже), и отрицательны, если распределение имеет длинный "хвост" слева от математического ожидания (пример — логистическое распределение). Поэтому функция моментов $\beta_1 = \mu_3/\mu_2^{3/2}$ часто служит *мерой асимметрии* и называется *коэффициентом асимметрии*.

Центральные моменты четных порядков всегда положительны, через них выражают *коэффициент эксцесса*, который характеризует остроту пика функции плотности вероятности и задается выражением $\beta_2 = \mu_4/\mu_2^2 - 3$. Для нормального распределения (см. ниже) $\beta_1 = 0$ и $\beta_2 = 0$. Распределения с положительным эксцессом обычно имеют более острый пик, чем график функции плотности нормального распределения, а распределения с отрицательным β_2 —

более сглаженный пик по сравнению с нормальным (например — распределение Стьюдента, которое описано ниже).

Другими характеристиками местоположения распределений могут служить *медиана* и *мода*. Медианой называют такое значение m , которое делит распределение на две равновероятные половины, т.е. $P(X < m) = P(X > m) = 1/2$. Отметим, что для дискретного распределения медиана не всегда вычисляется однозначно.

Мода μ определяется для непрерывных распределений, имеющих плотность вероятности, и соответствует такому значению случайной величины, которое является точкой максимума для функции плотности вероятностей. Обычно в статистике имеют дело с *одно модальными* распределениями, т.е. с такими, функция плотности вероятности которых имеет один максимум¹. Для симметричных одномодальных распределений математическое ожидание, мода и медиана совпадают. Отметим, что для большинства одномодальных распределений математическое ожидание, медиана и мода располагаются на числовой оси в том порядке, в котором они здесь перечислены, либо в обратном (это называется "алфавитное правило"). Таким образом, медиана лежит между математическим ожиданием и модой, причем ближе к математическому ожиданию. Для одномодальных распределений определена специальная мера асимметрии — *коэффициент асимметрии Пирсона*, который вычисляется по формуле $s = (MX - \mu)/\sigma$, где μ — мода, σ — корень из дисперсии. Для симметричных распределений коэффициент Пирсона равен нулю, он характеризует степень отклонения моды от математического ожидания.

В математической статистике также широко используются *квантили* случайных величин. Квантилью порядка p случайной величины X называется такое число ξ_p , что $P(X < \xi_p) = p$. Медиана является квантилью порядка $1/2$. Квантили некоторых порядков имеют специальные названия: *квартили* $\xi_{0.25}$, $\xi_{0.5}$, $\xi_{0.75}$, *децили* $\xi_{0.1}$, $\xi_{0.2}$, ..., $\xi_{0.9}$, *процентили* $\xi_{0.01}$, $\xi_{0.02}$, ..., $\xi_{0.99}$ делят область изменения случайной величины X соответственно на 4, 10 и 100 интервалов, значения из которых случайная величина X принимает с равными вероятностями. Для многих вероятностных распределений значения квантилей заданного уровня подсчитаны, сведены в специальные таблицы и используются при построении статистических критериев.

1.2.4. Вероятностные неравенства

В теории вероятностей и математической статистике большую роль играют неравенства, связывающие вероятности попадания случайной величины X в определенный интервал с числовыми характеристиками распределения. Наиболее общим неравенством такого типа является *неравенство Чебышева*, которое справедливо для любого вероятностного распределения случайной величины X :

$$P(|MX - X| \geq k\sigma) \leq 1/k^2.$$

Здесь и далее в этом разделе MX — математическое ожидание, $\sigma^2 = DX$ — дисперсия случайной величины X . Предполагается, что $k > 0$. Если случайная величина X принимает только положительные значения, то имеет место *неравенство Маркова* $P(X \geq kMX) \leq 1/k$.

Для случайных величин, имеющих одномодальное распределение, доказано несколько подобных неравенств, которые в общем случае точнее, чем неравенство Чебышева.

¹ Для таких распределений также встречаются названия унимодальное и одновершинное.

Неравенство Гаусса:

$$P(|X - MX| \geq k\sigma) \leq \frac{4}{9} \frac{1 + s^2}{(k - |s|)^2}, \quad k > |s|,$$

здесь s — коэффициент асимметрии Пирсона (см. предыдущий раздел)². Если распределение симметрично (в этом случае $s = 0$), тогда неравенство Гаусса имеет вид (сравните его с неравенством Чебышева)

$$P(|X - MX| > k\sigma) < \frac{4}{9k^2}$$

Если в качестве меры асимметрии распределения использовать величину $\delta = v/\sigma$, где $v = M|X - \mu|$, тогда для одномодальных распределений справедливо *неравенство Пика*

$$P(|X - MX| \geq k\sigma) \leq \frac{4}{9} \frac{1 - \delta^2}{(k - \delta)^2},$$

которое иногда точнее неравенства Гаусса.

Примеры использования этих неравенств приведены в главе 2 при построении доверительных интервалов.

1.2.5. Двумерные распределения

Рассмотрим кратко двумерные случайные величины $Z = (X, Y)$. Вероятностные свойства таких случайных величин характеризуют функции совместного распределения $F(x, y)$, которые определяются так же, как для одномерных величин, т.е. $F(x, y) = P(X < x \text{ и } Y < y)$. Для каждой составляющей X и Y случайной величины Z существуют *частные функции распределения*:

$$F_1(x) \equiv F(X < x) \equiv P(X < x \text{ и } Y < \infty) = F(x, \infty),$$

$$F_2(y) \equiv F(Y < y) \equiv P(X < \infty \text{ и } Y < y) = F(\infty, y).$$

Отметим, что функция $F(x, y)$ полностью определяет функции $F_1(x)$ и $F_2(y)$. Однако эти функции определяют функцию $F(x, y)$ только в том случае, когда компоненты X и Y независимы; тогда $F(x, y) = F_1(x)F_2(y)$.

Можно вычислить любые моменты каждой составляющей X и Y (если, конечно, они существуют), например MX, DX, MY, DY . Можно также вычислить различные смешанные моменты случайных величин X и Y . Среди смешанных моментов выделяют *ковариацию* величин X и Y , определяемую как математическое ожидание от произведения $(X - MX)(Y - MY)$, т.е.

$$\text{cov}(X, Y) = M[(X - MX)(Y - MY)].$$

Нормированную на дисперсии ковариацию называют *коэффициентом корреляции* ρ случайных величин X и Y :

$$\rho = \frac{\text{cov}(X, Y)}{\sqrt{DX \cdot DY}}.$$

Существует другой вариант неравенства Гаусса, в котором рассматривается интервал, симметричный относительно моды (а не относительно математического ожидания). Для симметричных *одномодальных* распределений эти два варианта неравенства совпадают.

Значение этого коэффициента лежит между -1 и 1. Он характеризует степень линейной зависимости между величинами X и Y . Если X и Y связаны строго линейно (например, $Y = -2X + 5$), то абсолютная величина ρ равна 1, если X и Y независимы, то $\rho = 0$. Однако нулевая корреляция *не означает* независимость X и Y (за исключением случая, когда случайные величины X и Y имеют нормальное распределение) — из этого следует только отсутствие какой-либо линейной зависимости между X и Y .

1.3. Функции от случайных величин

Функция от случайных величин также является случайной величиной. В принципе, любую случайную величину можно представить в виде функции от некоторой другой случайной величины (например, как функцию от равномерно распределенной случайной величины; см. приведенную ниже теорему). Преобразование случайных величин широко применяется в статистическом анализе. Функции от случайных величин также используются при генерировании случайных величин.

Пусть случайные величины X и Y связаны взаимно однозначным соответствием $Y = \varphi(X)$ и $X = \psi(Y)$, где ψ — функция, обратная к функции φ . Обозначим через $f_X(x)$, $f_Y(y)$, $F_X(x)$ и $F_Y(y)$ плотности вероятностей и функции распределения случайных величин X и Y соответственно. Они связаны между собой следующими формулами:

$$f_Y(y) = f_X(\psi(y)) \cdot |\psi'(y)|, \quad F_Y(y) = F_X(\psi(y)).$$

Имеют место также "обратные" формулы (если поменять местами X и Y и функцию ψ заменить на φ), показывающие зависимость распределения случайной величины X от распределения величины Y .

В общем случае (если не требовать взаимно однозначного соответствия между X и Y) функция распределения $F_Y(y)$ связана с функцией $F_X(x)$ формулой

$$F_Y(y) = \int_{\varphi(x) \leq y} dF_X(x).$$

Отметим, что для моментов случайных величин X и Y нет простых формул, отображающих их зависимость. (Формула $MY = \varphi(MX)$, которую часто пытаются использовать студенты, в общем случае неверна.) Однако можно указать простую зависимость между квантилями ζ_p и ξ_p случайных величин X и Y соответственно: если φ — возрастающая функция, то $\zeta_p = \varphi(\xi_p)$ для любого p ($0 < p < 1$); если же φ — убывающая функция, то в этом случае $\zeta_p = \varphi(\xi_{1-p})$. Для медиан $\zeta_{0,5}$ и $\xi_{0,5}$ соотношение $\zeta_{0,5} = \varphi(\xi_{0,5})$ справедливо как для возрастающей, так и для убывающей функции φ .

Приведем две теоремы, которые находят применение в статистическом анализе.

Теорема. Пусть случайная величина X имеет непрерывную функцию распределения $F(x)$. Тогда случайная величина $Y = F(X)$ распределена равномерно на интервале $[0, 1]$.

Отсутствие свойства взаимно однозначного отображения $X \leftrightarrow Y$ изменяет (усложняет) приведенные ниже формулы, но не является принципиальным препятствием для их построения.

Теорема⁴. Пусть $G(x)$ — функция, обратная к непрерывной строго монотонной функции распределения $F(x)$. Тогда случайная величина $Y = G(X)$, где случайная величина X распределена равномерно на интервале $[0, 1]$, имеет функцию распределения $F(x)$.

Первая теорема используется для построения так называемых *пробит-графиков* на этапе предварительного анализа выборочных распределений (см. главу 9). Вторая теорема лежит в основе *метода обратных функций* генерирования случайных величин, активно применяемого на практике (см. главу 7).

1.3.1. Линейное преобразование случайных величин

Это простейшая зависимость вида $Y = aX + b$ между случайными величинами X и Y . В этом случае $F_Y(x) = F_X((x - b)/a)$. Если случайная величина X непрерывна (т.е. существует ее плотность вероятности $f_X(x)$), тогда $f_Y(x) = \frac{1}{|a|} f_X\left(\frac{x-b}{a}\right)$

Между моментами случайных величин X и Y существуют такие соотношения:

$$MY = M(aX + b) = MX + b, DY = D(aX + b) = a^2 DX,$$

$$M(Y)^r = M(aX + b)^r = a^r m_r + C_r^1 a^{r-1} b m_{r-1} + \dots + C_r^{r-1} a b^{r-1} m_1 + b^r$$

здесь $M(X)^k$ — начальные (относительно $x = 0$) моменты порядка k случайной величины X , $C_r^k = \frac{r!}{k!(r-k)!}$ — биномиальные коэффициенты.

Для статистического анализа особый интерес представляет линейное *преобразование к стандартному виду* (нормирование случайной величины). Если случайная величина Y имеет математическое ожидание MX и дисперсию σ^2 , тогда случайная величина $Y = \frac{X - MX}{\sigma}$, у которой $MY = 0$ и $DY = 1$, называется *стандартизованной* (нормированной) случайной величиной. Нормирование случайных величин часто применяется на предварительном этапе статистического анализа (см. главу 8).

1.3.2. Суммы случайных величин

Для случайной величины $Z = X + Y$ всегда верно (вне зависимости от того, будут ли случайные величины X и Y независимыми), что

$$MZ = M(X + Y) = MX + MY.$$

Дисперсия случайной величины Z вычисляется по формуле

$$DZ = D(X + Y) = DX + DY + 2\text{cov}(X, Y).$$

Если случайные величины X и Y независимы, то $D(X + Y) = DX + DY$.

⁴ В формулировках теорем мы намерено наложили жесткие ограничения (непрерывность и строгую монотонность) на функцию распределения $F(x)$, чтобы избежать проблем с неоднозначностью обратной функции $G(x)$ в случае разрывной или нестрого монотонной функции $F(x)$. На практике эти теоремы используются для любых функций распределения, если доопределить их должным образом.

В случае, когда случайные величины X и Y имеют совместную плотность вероятности $f(x, y)$, тогда плотность вероятности $g(x)$ случайной величины $Z = X + Y$ выражается формулой

$$S(z) = \int f(x, z-x) dx = \int f(z-y, y) dy .$$

В частности, когда случайные величины X и Y независимы (в этом случае $f(x, y) = Mx) My)$), тогда

$$g(z) = \int f_x(x) f_y(z-x) dx = \int f_x(x) f_y(z-x) dx .$$

N

Если Z является суммой N случайных величин X_1, X_2, \dots, X_N , т.е. $Z = \sum_{i=1}^N X_i$, тогда $MZ = \sum_{i=1}^N MX_i$. Дисперсия суммы случайных величин вычисляется по формуле

$$DZ = \sum_{i=1}^N DX_i + 2 \sum_{i < j} \text{Cov}(X_i, X_j) .$$

Таким образом, дисперсия суммы случайных величин равняется сумме их дисперсий и суммы ковариаций всех возможных пар случайных величин. Для независимых случайных величин X_1, X_2, \dots, X_N $DZ = \sum_{i=1}^N DX_i$.

Если количество слагаемых в сумме $Z = \sum_{i=1}^N X_i$ неограниченно возрастает, то при достаточно общих условиях, накладываемых на случайные величины X_i , распределение случайной величины Z сходится к нормальному распределению. Перечисление этих условий составляет содержание *центральной предельной теоремы* теории вероятностей.

1.3.3. Центральная предельная теорема

Исключительное значение центральных предельных теорем объясняется тем, что они являются теоретической основой применения нормального распределения при решении многих практических задач. Всегда, когда можно предположить, что рассматриваемая величина является суммой большого числа случайных факторов, влияние каждого из которых пренебрежимо мало, ее распределение будет близко к нормальному распределению. Такими величинами являются, например, ошибки регистрации в измерительных приборах, результаты случайного эксперимента, зависящего от многих малых факторов, рассеивание электронов при бомбардировке ими мишеней и т.д.

Приведем простейший вариант центральной предельной теоремы, относящийся к суммам независимых одинаково распределенных слагаемых с конечной дисперсией. Именно этот вариант теоремы служит основой для построения различных асимптотических оценок выборочных параметров распределений в статистическом анализе (см. раздел 2.2).

Теорема. Пусть $X_1, X_2, \dots, X_n, \dots$ — последовательность независимых одинаково распределенных случайных величин с математическим ожиданием $MX_k = m$ и конечной дисперсией $DX_k = \sigma^2 > 0$. Обозначим $Y_n = X_1 + X_2 + \dots + X_n$. Тогда при $n \rightarrow \infty$ для любого x

$$P\left(\frac{Y_n - nm}{\sigma\sqrt{n}} < x\right) \rightarrow \Phi(x),$$

где $\Phi(x)$ — функция распределения стандартного нормального закона.

Последовательность случайных величин Y_n называется *асимптотически нормальной*.

Существуют более общие варианты центральной предельной теоремы, справедливые для последовательностей $Y_n = X_1 + X_2 + \dots + X_n$, когда X_k могут иметь различные распределения и быть зависимыми. /Различные варианты теоремы можно найти в [6].

1.4. Примеры дискретных распределений

Приведем примеры распределений, которые часто встречаются при проведении статистического анализа.

1.4.1. Равномерное дискретное распределение

Случайная величина X имеет равномерное дискретное распределение, если она принимает конечное число различных значений с одинаковой вероятностью. Пусть, для определенности, величина X может принимать значения $1, 2, \dots, n$. Тогда $P(X = i) = 1/n$ для всех целых значений i из интервала $[1, n]$. Отметим, что в этом случае $MX = (n + 1)/2$, $DX = (n + 1)(2n + 1)/6$. График этого распределения для $n = 10$ показан на рис. 1.4.

Это распределение часто используется для моделирования равновероятных дискретных событий.

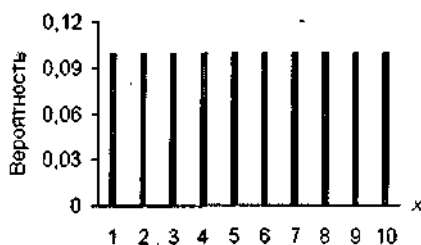


Рис. 1.4. Дискретное равномерное распределение

1.4.2. Распределение Бернулли

Случайная величина X имеет распределение Бернулли с параметром p ($0 < p < 1$), если $P(X = 1) = p$ и $P(X = 0) = 1 - p$. Таким образом, случайная ве-

личина X может принимать только два значения, 1 и 0, с вероятностями p и $1 - p$ соответственно. Отметим, что $MX = p$ и $DX = p(1 - p)$.

Это распределение играет фундаментальную роль в теории вероятностей и математической статистики, поскольку является моделью любого случайного эксперимента, результатом которого может быть один из двух возможных исходов: исход "1" произойдет с вероятностью p и исход "0" — с вероятностью $1 - p$ (исход "1" часто называют "успехом", а исход "0" — "неудачей").

1.4.3. Биномиальное распределение

Случайная величина X имеет биномиальное распределение с параметрами n и p ($0 < p < 1$, $n \geq 1$), если

$$P(X = k) = C_n^k p^k (1 - p)^{n-k}$$

Здесь $C_n^k = \frac{n!}{k!(n-k)!}$ — биномиальный коэффициент. Для этой случайной ве-

личины $MX = np$, $DX = np(1 - p)$. Распределение вероятностей для значений параметров $n = 20$ и $p = 0,5$ показано на рис. 1.5.

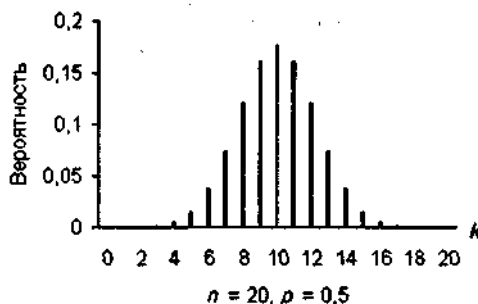


Рис. 1.5. Биномиальное распределение

Биномиальное распределение является моделью случайных экспериментов, состоящих из n независимых одинаковых испытаний. В результате каждого из них с вероятностью p может произойти исход "1" и с вероятностью $1 - p$ — исход "0". Принятым названием для такой модели случайных экспериментов является *схема Бернулли*. Случайная величина, равная количеству k исходов "1" в n испытаниях, имеет биномиальное распределение. Для вычисления вероятностей $P(X = k)$ при достаточно больших n и при условии, что $1/(n+1) < p < n/(n+1)$, часто используются приближенные формулы, основанные на аппроксимации этого распределения нормальным. В Excel есть функция БИНОМРАСП (см. главу 4), которая позволяет вычислять как значения вероятностей $P(X = k)$ при любых n , p и k , так и значения функции распределения $F(x)$. (Графики на рис. 1.5 построены с помощью этой функции.)

1.4.4. Распределение Пуассона

Случайная величина X имеет распределение Пуассону с параметром X ($X > 0$), если

$$P(X = k) = e^{-\lambda} \frac{\lambda^k}{k!}, \quad k = 0, 1, 2, \dots$$

Для этого распределения математическое ожидание и дисперсия совпадают, т.е. $MX = DX = X$. Распределения вероятностей для двух значений X показаны на рис. 1.6.

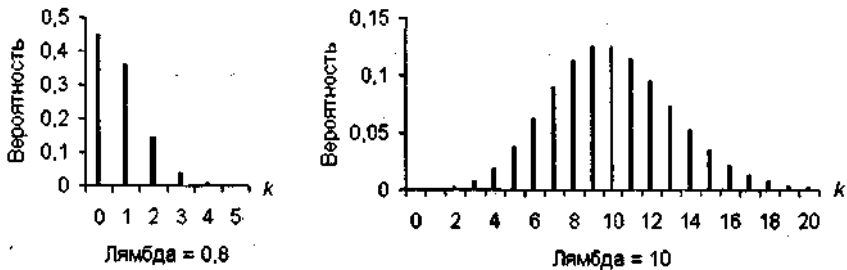


Рис. 1.6. Распределение Пуассона

Распределение Пуассона играет важную роль в теории вероятностей и математической статистике. Оно является моделью для описания случайного числа появлений определенных событий в фиксированный промежуток времени или в фиксированной области пространства. Традиционными примерами случайных величин, подчиняющихся распределению Пуассона, являются число альфа-частиц, испускаемых радиоактивным источником за определенный промежуток времени; количество бактерий, видимых под микроскопом; мутации, вызванные радиацией; количество звезд в определенной области звездного неба; количество деревьев на участке леса и т.д. В Excel для вычисления вероятностей $P(X = k)$ и значений функции распределения $F(x)$ есть функция ПУАССОН (см. главу 4).

Отметим также соотношения между распределениями Пуассона и χ^2 (см. раздел 1.5.5), которые используются при построении интервальных оценок для параметра X (см. раздел 2.3.8): $P(X \geq k) = P(Y \leq 2\lambda)$, где Y — случайная величина, имеющая χ^2 -распределение с $2k$ степенями свободы, и $P(X < k) = P(Z > 2X)$, где Z — случайная величина, имеющая χ^2 -распределение с $2(k + 1)$ степенями свободы.

1.4.5. Геометрическое распределение

Случайная величина X имеет геометрическое распределение с параметром p ($0 < p < 1$), если

$$P(X = k) = p(1 - p)^k, \quad k = 0, 1, 2, \dots$$

Для этой случайной величины $MX = (1 - p)/p$, $DX = (1 - p)/p^2$. Распределение вероятностей для значения параметра $p = 0,7$ показано на рис. 1.7.

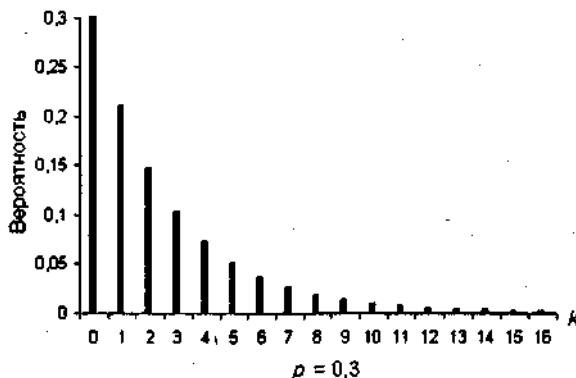


Рис. 1.7. Геометрическое распределение

Это распределение является частным случаем отрицательного биномиального распределения при параметре $r = 1$ (см. раздел 1.4.7) и описывает число испытаний в схеме Бернулли (раздел 1.4.3), необходимых для того, чтобы получить исход "1" ровно один раз.

1.4.6. Гипергеометрическое распределение

Случайная величина X имеет гипергеометрическое распределение с параметрами N , n и p ($N \geq n \geq 0$, $0 < p < 1$), если

$$P(X = k) = \frac{C_{np}^k C_{N(1-p)}^{n-k}}{C_N^n}, \quad k = 0, 1, 2, \dots, n.$$

Здесь C_n^k — биномиальный коэффициент. Для этой случайной величины $MX = np$, $DX = np(1-p) \frac{N-n}{N-1}$. Распределение вероятностей для значений параметров $N \leq 100$, $n = 10$ и $p = 0.4$ показано на рис. 1.8.

Типичная ситуация, в которой появляется гипергеометрическое распределение, следующая: проверяется партия готовой продукции объемом N , в которой любое изделие с вероятностью p является годным и, соответственно, с вероятностью $1 - p$ — бракованным. Случайным образом выбираются n изделий. Гипергеометрическое распределение описывает число годных изделий среди n выбранных изделий.

Если $n/N < 0.1$, это распределение хорошо аппроксимируется биномиальным распределением. В Excel имеется функция ГИПЕРГЕОМЕТ, вычисляющая вероятность $P(X = K)$ при заданных значениях N , n , p и k (см. раздел 4.6.6).

1.4.7. Отрицательное биномиальное распределение (распределение Паскаля)

Случайная величина X имеет отрицательное биномиальное распределение (распределение Паскаля) с параметрами g и p ($0 < p < 1$), если

$$P(X = k) = C_{g+k-1}^{k-1} p^k (1-p)^{g-k}, \quad k = 0, 1, 2, \dots$$

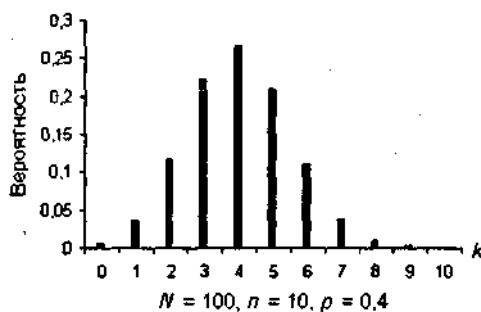


Рис. 1.8. Гипергеометрическое распределение

Здесь C^* — биномиальный коэффициент. Для этой случайной величины $MX = r(1 - p)/p$, $DX = r(1 - p)/p^2$. Распределение вероятностей для значений параметров $r = 10$ и $p = 0,8$ показано на рис. 1.9.

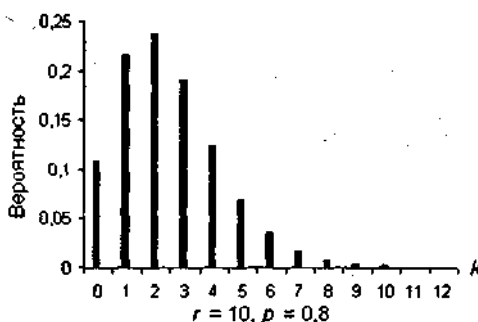


Рис. 1.9. Отрицательное биномиальное распределение

При натуральном r отрицательное биномиальное распределение описывает число испытаний в схеме Бернулли, необходимых для того, чтобы получить исход "I" ровно r раз. Это распределение часто появляется в популяционной биологии.

В Excel имеется функция ОТРБИНОМРАСП, вычисляющая вероятности $P(X = k)$ при заданных значениях r , p и k (см. раздел 4.6.10).

1.5. Примеры непрерывных распределений

1.5.1. Равномерное непрерывное распределение

Случайная величина X имеет равномерное распределение на интервале $[a, b]$, если ее плотность вероятности (рис. 1.10) вычисляется по формуле

$$f(x) = \begin{cases} \frac{1}{b-a}, & \text{если } x \in [a, b], \\ 0, & \text{если } x \notin [a, b]. \end{cases}$$

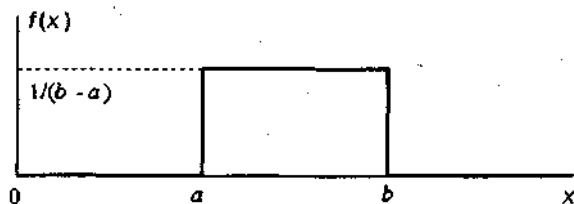


Рис. 1.10. Плотность равномерного распределения

Для этой случайной величины $MX = (a + b)/2$, $DX = (b - a)^2/12$, $\beta_1 = 0$, $\beta_2 = -1,2$. Случайная величина $Y = (X - a)/(b - a)$ распределена равномерно на интервале $[0, 1]$. Равномерное распределение является непрерывным аналогом дискретного равномерного распределения, описывающего случайные эксперименты с равновероятными исходами.

Теоремы из раздела 1.3, показывающие взаимосвязь равномерного распределения с другими типами распределений, объясняют широкое использование равномерного распределения в статистическом моделировании (более подробно об этом речь идет в главе 7). В Excel функция СЛЧИС генерирует случайные числа, равномерно распределенные на интервале $[0, 1]$ (см. раздел 4.13.1).

1.5.2. Треугольное распределение

Случайная величина X имеет треугольное распределение (называемое также распределением Симпсона) на интервале $[a, b]$, если ее плотность вероятности (рис. 1.11) вычисляется по формуле

$$f(x) = \begin{cases} \frac{2}{b-a} - \frac{2}{(b-a)^2} |a+b-2x|, & \text{если } x \in [a, b], \\ 0, & \text{если } x \notin [a, b]. \end{cases}$$

Для этой случайной величины $MX = \frac{1}{3}(b(a^2 + b^2) - (a+b)^2)$, $DX = (b - a)^3/24$. Если X_1 и X_2 — независимые случайные величины, равномерно распределенные на интервале $\left[\frac{a}{2}, \frac{b}{2}\right]$, то случайная величина $X = X_1 + X_2$ имеет треугольное распределение на интервале $[a, b]$.

1.5.3. Показательное (экспоненциальное) распределение

Случайная величина X имеет показательное (экспоненциальное) распределение с параметром λ ($\lambda > 0$), если ее плотность вероятности (рис. 1.12) вычисляется по формуле

$$f(x) = \begin{cases} \lambda e^{-\lambda x}, & \text{если } x \geq 0, \\ 0, & \text{если } x < 0. \end{cases}$$

Для этой случайной величины $MX = 1/\lambda$, $DX = 1/\lambda^2$; ее функция распределения вычисляется по простой формуле $F(u) = 1 - e^{-\lambda u}$ ($u \geq 0$). Это распределение

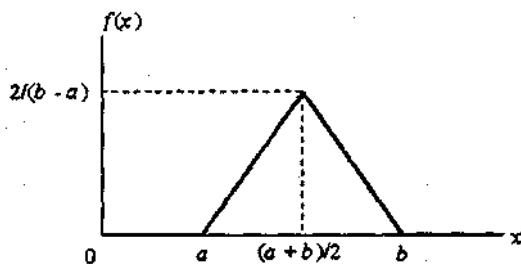


Рис. 1.11. Плотность треугольного распределения

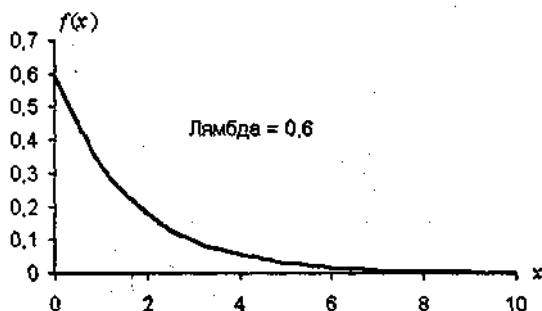


Рис. 1.12. Плотность показательного распределения

часто встречается в моделировании случайных процессов (оно обладает так называемым свойством отсутствия последствия). В Excel функция ЭКСПРАСП вычисляет значения плотности и функции распределения (см. раздел 4.6.15).

1.5.4. Нормальное распределение

Случайная величина X имеет нормальное распределение с параметрами m и σ^2 , если ее плотность вероятности (рис. 1.13) вычисляется по формуле

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-m)^2}{2\sigma^2}}, \quad x \in (-\infty, \infty).$$

Для этой случайной величины $MX = m$, $DX = \sigma^2$, $\beta_1 = 0$, $\beta_2 = 0$. Нормальное распределение называют также гауссовским распределением, законом Гаусса, вторым законом Лапласа, распределением Гаусса-Лапласа и др.

Если $m = 0$ и $\sigma^2 = 1$, то распределение называется *стандартным* нормальным распределением. Линейное преобразование $Y = (X - m)/\sigma$ приводит произвольную нормально распределенную величину X к стандартному нормальному распределению.

Фундаментальная роль, которую играет нормальное распределение в теории вероятностей и математической статистике, объясняется тем, что при доста-

точно широких условиях распределение суммы случайных величин с ростом числа слагаемых асимптотически сходится к нормальному. Соответствующие условия сходимости приведены в *центральной предельной теореме* теории вероятностей (см. раздел 1.3.3).

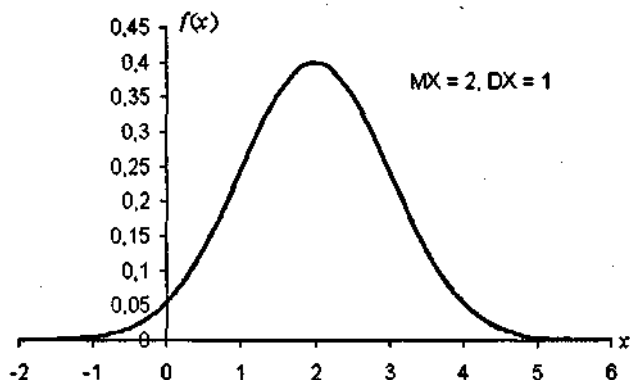


Рис. 1.13. Плотность нормального распределения

Нормально распределенная случайная величина с большой вероятностью принимает значения, близкие к своему математическому ожиданию. Это свойство нормального распределения формулируется как *правило сигм*:

$$P(|X - m| \geq k\sigma) = \begin{cases} 0,3173..., & k = 1, \\ 0,0455..., & k = 2, \\ 0,0027..., & k = 3. \end{cases}$$

Чаще всего используют *правило трех сигм*, которое находит широкое применение в математической статистике при построении доверительных интервалов.

В Excel функции НОРМСТРАСП и НОРМРАСП (см. разделы 4.6.8 и 4.6.9) вычисляют значения плотности вероятности и функции распределения соответственно стандартного и произвольного нормального распределений, а функции НОРМСТОБР и НОРМОБР — значения функций, обратных к функциям распределения стандартного и произвольного нормального законов (см. разделы 4.7.5 и 4.7.6). Последние функции можно использовать для генерирования нормально распределенных случайных величин (см. главу 7).

1.5.5. Распределение "хи-квадрат"

Случайная величина X имеет распределение χ^2 с n степенями свободы, если ее плотность вероятности (рис. 1.14) вычисляется по формуле

$$f(x) = \begin{cases} \frac{1}{2^{\frac{n}{2}} \Gamma(\frac{n}{2})} x^{\frac{n}{2}-1} e^{-\frac{x}{2}}, & \text{если } x \geq 0, \\ 0, & \text{если } x < 0. \end{cases}$$

Здесь и далее $\Gamma(x)$ — гамма-функция Эйлера⁵. Для данного распределения $MX = n$, $DX = 2n$, $\beta_1 = 2\sqrt{\frac{2}{n}}$, $\beta_2 = 12/n$. При $n \geq 2$ мода находится в точке $x = n - 2$.

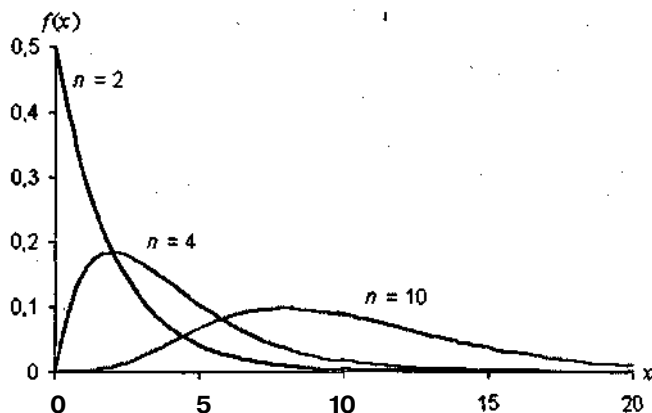


Рис. 1.14. Распределение у?

Многочисленные применения этого распределения в теории вероятностей основаны на том факте, что если X_1, X_2, \dots, X_n — независимые случайные величины, имеющие стандартное нормальное распределение, то случайная величина $Y = \sum_{i=1}^n X_i^2$ имеет распределение χ^2 с n степенями свободы. В математической статистике распределение χ^2 применяется при построении целого ряда разнообразных критериев, в том числе при согласовании выборочных данных с выбранным законом распределения и в методе наименьших квадратов (см. главы 2 и 3).

В Excel есть три функции, ХИ2РАСП, ХИ2БР и ХИ2ТЕСТ, связанные с распределением χ^2 . Подробно эти функции описаны в главе 4.

1.5.6. Распределение Стьюдента

Случайная величина X имеет распределение Стьюдента (t -распределение) с n степенями свободы, если ее плотность вероятности (рис. 1.15) вычисляется по формуле

$$f(x) = \frac{\Gamma\left(\frac{n+1}{2}\right)}{\sqrt{n\pi}\Gamma\left(\frac{n}{2}\right)} \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}}, \quad x \in (-\infty, \infty).$$

⁵ Значения гамма-функции $\Gamma(n) = \int_0^\infty e^{-x} x^{n-1} dx$ можно получить с помощью функции Excel ГАММАПРОГ, вычисляющей натуральный логарифм гамма-функции. Также отметим, что $\Gamma(n) = (n-1)!$, если n — натуральное число.

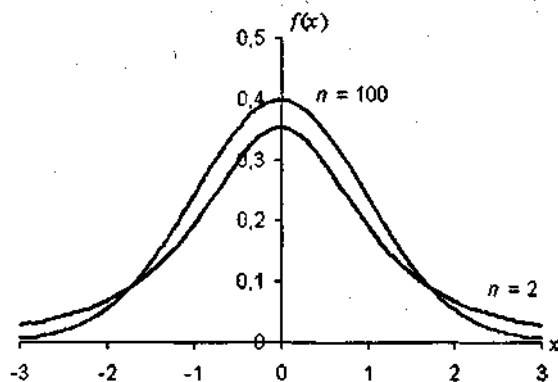


Рис. 1.15. Плотность распределения Стьюдента

Для этого распределения при $n > 2$ $MX = 0$, $DX = n/(n - 2)$ (если $n \leq 2$, то $DX = \infty$), $\beta_1 = 0$, $\beta_2 = 6/(n - 4)$ (при $n > 4$). При больших значениях n распределение Стьюдента асимптотически приближается со стандартным нормальным распределением.

Распределение Стьюдента имеет многочисленные применения в математической статистике. Если Y — случайная величина, имеющая стандартное нормальное распределение, а Z — случайная величина, имеющая распределение χ^2 с n степенями

свободы, тогда случайная величина $X = Y / \sqrt{Z/n}$ имеет t -распределение также с n степенями свободы. (О применении распределения Стьюдента речь идет в главе 2.)

В Excel имеются функции СТЬЮДРАСП и СТЬЮДОБР, вычисляющие соответственно значения функции распределения и обратной к ней функции (см. разделы 4.6.12 и 4.7.7).

1.5.7. F -распределение

Случайная величина X имеет F -распределение (распределение Снедекора) с (m, n) степенями свободы ($m, n \geq 1$), если ее плотность вероятности (рис. 1.16) вычисляется по формуле

$$f(x) = \begin{cases} \frac{\Gamma\left(\frac{m+n}{2}\right)}{\Gamma\left(\frac{m}{2}\right)\Gamma\left(\frac{n}{2}\right)} \left(\frac{m}{n}\right)^{\frac{m}{2}} x^{\frac{m}{2}-1} \left(1 + \frac{m}{n}x\right)^{-\frac{m+n}{2}}, & \text{если } x \geq 0, \\ 0, & \text{если } x < 0. \end{cases}$$

Для этого распределения $MX = \frac{n}{n-2}$ (при $n > 2$), $DX = \frac{2n^2(m+n-2)}{m(n-2)^2(n-4)}$ (если $n > 4$). При $m \geq 2$ мода находится в точке $x = \frac{n(m-2)}{m(n+2)}$.

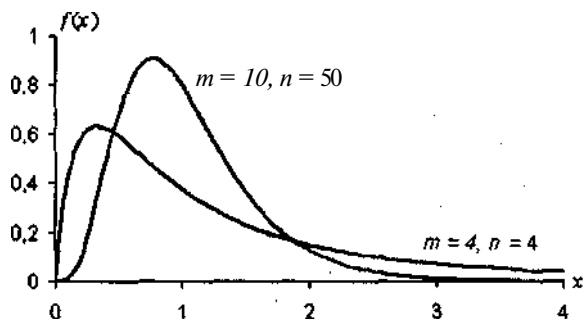


Рис. 1.16. Плотность F-распределения

Если независимые случайные величины Y_1 и Y_2 имеют распределение y с m и n степенями свободы соответственно, тогда случайная величина $X = \frac{Y_1/m}{Y_2/n}$ будет иметь F-распределение.

F-распределение играет основную роль при сравнении выборочных дисперсий из нормально распределенных совокупностей. Оно также широко используется в регрессионном и дисперсионном анализе. В Excel имеются функции FРАСн и FРАСгОВР, которые вычисляют значения соответственно функции распределения и обратной к ней функции (см. разделы 4.6.1 и 4.7.1).

1.5.8. Логарифмически нормальное распределение

Случайная величина X имеет логарифмически нормальное (логнормальное) распределение с параметрами m и σ^2 , если ее плотность вероятности (рис. 1.17) вычисляется по формуле

$$f(x) = \begin{cases} \frac{1}{x\sigma\sqrt{2\pi}} \exp\left\{-\frac{(\ln(x-m))^2}{2\sigma^2}\right\}, & \text{если } x > 0, \\ 0, & \text{если } x \leq 0. \end{cases}$$

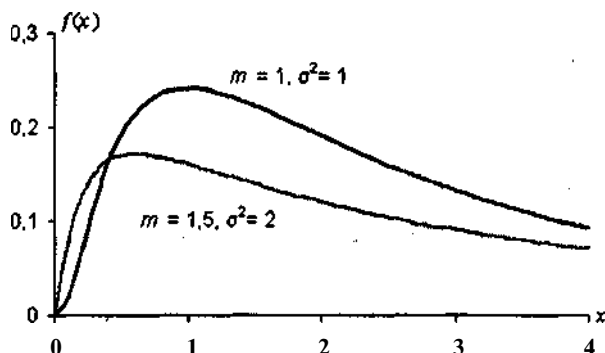


Рис. 1.17. Плотность логнормального распределения

Для этой случайной величины $MX = \exp(m + \sigma^2/2)$, $DX = (\exp(\sigma^2) - 1)\exp(2m + \sigma^2)$. Если случайная величина X имеет логнормальное распределение, то ее логарифм $Y = \ln X$ распределен по нормальному закону с математическим ожиданием m и дисперсией σ^2 .

Это распределение находит применение в теории надежности, статистической физике, экономической статистике, биологии и т.д. В Excel имеются функции ЛОГНОРМРАСП и ЛОГНОРМОБР, которые вычисляют значения соответственно функции распределения и обратной к ней функции (см. разделы 4.6.7 и 4.7.4).

1.5.9. Бета-распределение

Случайная величина X имеет бета-распределение с параметрами α и β ($\alpha > 0$, $\beta > 0$), если ее плотность вероятности (рис. 1.18) вычисляется по формуле

$$f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, & \text{если } x \in [0,1], \\ 0, & \text{если } x \notin [0,1]. \end{cases}$$

Для этой случайной величины $MX = \alpha/(\alpha + \beta)$, $DX = \alpha\beta/((\alpha + \beta)^2(\alpha + \beta + 1))$. Если $\alpha > 1$ и $\beta > 1$, то распределение одномодально с модой в точке $x = (\alpha - 1)/(\alpha + \beta - 1)$. При $\alpha = \beta = 1$ бета-распределение является равномерным на интервале $[0, 1]$ распределением, при $\alpha = \beta = 2$ — треугольным, в случае $\alpha = \beta = 1/2$ оно называется *распределением арксинуса*, а при $\beta = \alpha + 1$ — *обобщенным распределением арксинуса*.

В математической статистике бета-распределение наиболее часто встречается в качестве распределения порядковых статистик (см. главу 2). В Excel функции БЕТАРАСП и БЕТАОБР вычисляют значения соответственно функции распределения и обратной к ней функции (см. разделы 4.6.2 и 4.7.2).

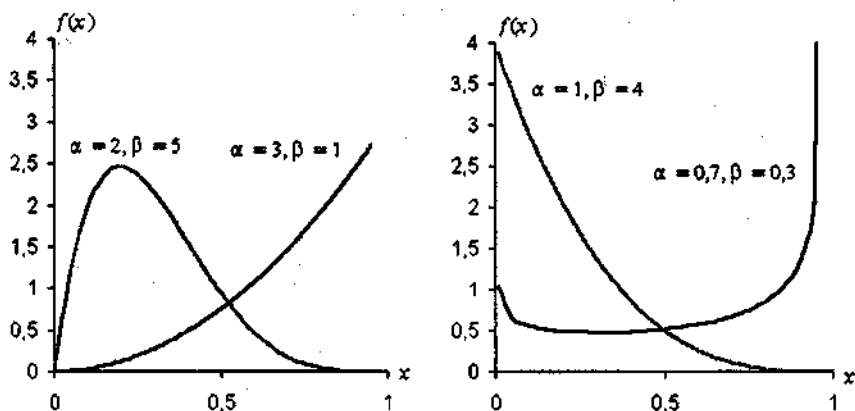


Рис. 1.18. Плотность бета-распределения

1.5.10. Гамма-распределение

Случайная величина X имеет гамма-распределение с параметрами α и λ ($\alpha > 0, \lambda > 0$), если ее плотность вероятности (рис. 1.19) вычисляется по формуле

$$f(x) = \begin{cases} \frac{\lambda^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\lambda x}, & \text{если } x > 0, \\ 0, & \text{если } x \leq 0. \end{cases}$$

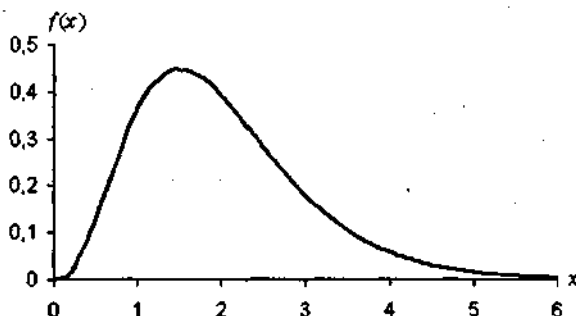


Рис. 1.19. Плотность гамма-распределения при $\alpha = 1$ и $\lambda = 0,5$

Для этой случайной величины $MX = \alpha/\lambda$, $DX = \alpha/\lambda^2$. При $\alpha \leq 1$ мода распределения находится в нуле, а при $\alpha \geq 1$ — в точке $x = (\alpha - 1)/\lambda$. Если $\alpha = 1$, то гамма-распределение совпадает с показательным распределением, а при $\alpha = n/2$, $\lambda = 1/2$ — с распределением χ^2 с n степенями свободы. В случае $\lambda = n$ и $\alpha = n$ (n — натуральное число) это распределение называют *распределением Эрланга* с параметрами n и μ . При натуральном α и $\lambda = 1$ гамма-распределение называется *показательно-степенным*.

Данное распределение и его частные случаи широко используются в теории вероятностей и математической статистике. В Excel функции ГАММАРАСП и ГАММАОБР вычисляют значения соответственно функции распределения и обратной к ней функции (см. разделы 4.6.5 и 4.7.3).

1.5.11. Распределение Вейбулла–Гнеденко

Случайная величина X имеет распределение Вейбулла–Гнеденко с параметрами α и λ ($\lambda > 0$), если ее плотность вероятности (рис. 1.20) вычисляется по формуле

$$f(x) = \begin{cases} \lambda \alpha x^{\alpha-1} e^{-\lambda x^\alpha}, & \text{если } x > 0, \\ 0, & \text{если } x \leq 0. \end{cases}$$

Для этой случайной величины

$$MX = \frac{\lambda^{-1/\alpha}}{\alpha} \Gamma\left(\frac{1}{\alpha}\right) \text{ и } DX = \lambda^{-2/\alpha} \left\{ \frac{2}{\alpha} \Gamma\left(\frac{2}{\alpha}\right) - \frac{1}{\alpha^2} \left[\Gamma\left(\frac{1}{\alpha}\right) \right]^2 \right\}.$$

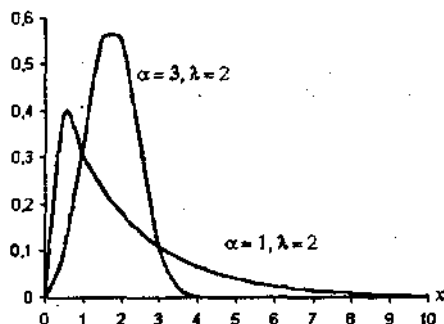


Рис. 1.20. Плотность распределения Вейбулла-Гнеденко при $\alpha=1$ и $\alpha=3$ и $\lambda=2$

Рис. 1.20. Плотность распределения Вейбулла-Гнеденко при $a=1$ и $a=3$ и $\lambda=2$

Распределение Вейбулла-Гнеденко часто используется в теории надежности, в частности для описания времени безотказной работы приборов. В Excel функция ВЕЙБУЛЛ вычисляет значения плотности вероятности и функции распределения (см. раздел 4.6.4).

1.5.12. Распределения Пирсона

Система распределений Пирсона основана на том, что плотности вероятности многих известных распределений подчиняются дифференциальному уравнению одного определенного типа, которое зависит от четырех параметров. В зависимости от значений этих параметров различают 12 типов распределений, среди которых такие распределения, как нормальное, гамма-распределение, бета-распределение, распределение Стюдента и другие. Уже это перечисление распределений, играющих фундаментальную роль в теории вероятностей и математической статистике, показывает важность системы распределений Пирсона. Поскольку в математической литературе полное описание распределений Пирсона встречается редко (обычно указываются только некоторые типы распределений), приведем их подробную классификацию. (Приведенный ниже материал, с небольшими дополнениями автора, взят из [8]. Другую классификацию кривых Пирсона можно найти в [4].)

Распределениями Пирсона называются непрерывные распределения, плотности вероятности которых являются решениями дифференциального уравнения

$$\frac{df(x)}{dx} = \frac{x+a}{b_0+2b_1x+b_2x^2} f(x),$$

где a, b_0, b_1, b_2 — параметры распределения. Эти параметры полностью определяются первыми четырьмя центральными моментами распределения. Пусть μ_k — k -й центральный момент, тогда

$$\begin{aligned} a &= \frac{\mu_3(\mu_4+3\mu_2^2)}{A}, & b_0 &= -\frac{\mu_2(4\mu_2\mu_4-3\mu_2^2)}{A}, \\ b_1 &= -\frac{\mu_2(\mu_4+3\mu_2^2)}{2A}, & b_2 &= -\frac{2\mu_2\mu_4-3\mu_2^2-6\mu_3^2}{A}, \end{aligned}$$

где $A = 10\mu_2\mu_4 - 18\mu_2^3 - 12\mu_3^2$.

Типы распределений Пирсона различают в соответствии со значениями корней квадратного уравнения $b_0 + 2b_1x + b_2x^2 = 0$. Введем обозначения: $D = b_0b_2 - b_1^2$, $\lambda = b_1^2/b_0b_2 = 1 - D$. Отметим, что большинство приведенных ниже формул для плотностей вероятности упрощается, если за начало отсчета взять моду распределения или математическое ожидание.

Тип I. $D < 0$, $\lambda < 0$ и $b_0 + 2b_1x + b_2x^2 = b_2(x + \alpha)(x - \beta)$, $\alpha, \beta > 0$. Обозначим как $m = (\alpha - a)/b_2(\alpha + \beta)$, $n = (\beta - a)/b_2(\alpha + \beta)$. Плотность вероятности этого типа распределений Пирсона определяется формулой

$$f(x) = \begin{cases} \frac{\alpha^{2m}\beta^{2n}}{(\alpha + \beta)^{m+n+1} B(m+1, n+1)} (\alpha + x)^m (\beta - x)^n, & \text{если } x \in [-\alpha, \beta], \\ 0, & \text{если } x \notin [-\alpha, \beta]. \end{cases}$$

Здесь и далее $B(m, n) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$ ($m > 0, n > 0$) — бета-функция.

Распределениями этого типа являются бета-распределения.

Тип II. $D < 0$, $\lambda = 0$ и $b_0 + 2b_1x + b_2x^2 = b_2(x^2 - \alpha^2)$, $\alpha = \sqrt{-b_0/b_2} > 0$. Обозначим как $m = 1/2b_2$. Плотность вероятности этого типа распределений определяется формулой

$$f(x) = \begin{cases} \frac{1}{\alpha^{2m+1} B(m+1, 1/2)} (\alpha^2 - x^2)^m, & \text{если } x \in [-\alpha, \alpha], \\ 0, & \text{если } x \notin [-\alpha, \alpha]. \end{cases}$$

Это распределение симметрично относительно точки $x = 0$.

Тип III. $D < 0$, $\lambda = \infty$ и $b_0 + 2b_1x + b_2x^2 = 2b_1(x + \alpha)$, $\alpha = b_0/2b_1$. Обозначим как $m = (a - \alpha)/2b_1$, $k = -1/2b_1$ ($k > 0$). Плотность вероятности этого типа распределений определяется формулой

$$f(x) = \begin{cases} \frac{k^{m+1}}{\Gamma(m+1)} (x + \alpha)^m e^{-k(x+\alpha)}, & \text{если } x > -\alpha, \\ 0, & \text{если } x \leq -\alpha. \end{cases}$$

Этот тип распределения является гамма-распределением.

Тип IV. $D > 0$, $0 < \lambda < 1$ и $b_0 + 2b_1x + b_2x^2 = b_2(x_1^2 + \alpha^2)$, $\alpha^2 = D/(b_2)^2$. Обозначим как $m = -1/2b_2 \geq 1/2$, $k = -b_1(2b_2 + 1)/(b_2)^2$. Плотность вероятности этого типа распределений определяется формулой

$$f(x) = c(\alpha^2 + x^2)^{-m} e^{-k \arctan(x/\alpha)}, \quad x \in (-\infty, \infty),$$

где $c^{-1} = \int_{-\infty}^{\infty} (\alpha^2 + x^2)^{-m} e^{-k \arctan(x/\alpha)} dx$.

Тип V. $D = 0$, $\lambda = 1$ и $b_0 + 2b_1x + b_2x^2 = b_2(x_1^2 + \alpha^2)$, $\alpha = b_1/b_2$. Обозначим как $m = -1/b_2 \geq 1$, $k = -b_1(2b_2 + 1)/b_2 > 0$. Плотность вероятности этого типа распределений определяется формулой

$$f(x) = \begin{cases} \frac{k^{m-1}}{\Gamma(m-1)} x^{-m} e^{-\frac{k}{x}}, & \text{если } x > 0, \\ 0, & \text{если } x \leq 0. \end{cases}$$

Тип VI. $D < 0$, $\lambda > 1$ и $b_0 + 2b_1x + b_2x^2 = b_2(x + \alpha)(x - \beta)$. Обозначим как $m = (\alpha - a)/b_2(\alpha + \beta) > 1$, $n = (\beta - a)/b_2(\alpha + \beta) > -1$. Плотность вероятности этого типа распределений определяется формулой

$$f(x) = \begin{cases} \frac{(\alpha + \beta)^{-(m+n+1)}}{B(-m-n-1, n+1)} (x + \alpha)^m (x - \beta)^n, & \text{если } x > \beta, \\ 0, & \text{если } x \leq \beta. \end{cases}$$

Тип VII. $D > 0$, $\lambda = 0$ и $b_0 + 2b_1x + b_2x^2 = b_2(x^2 + \alpha^2)$, $\alpha^2 = b_0/b_2$. Обозначим как $m = 1/2b_2 \geq 1/2$. Плотность вероятности этого типа распределений определяется формулой

$$f(x) = \frac{\alpha}{B\left(m - \frac{1}{2}, \frac{1}{2}\right)} (\alpha^2 + x^2)^{-m}, \quad \text{если } x \in (-\infty, \infty).$$

Распределение этого типа является распределением Стьюдента.

Тип VIII. $D < 0$, $\lambda < 0$ и $b_0 + 2b_1x + b_2x^2 = b_2(x + \alpha)x$, $\alpha = 2b_1/b_2$. Обозначим как $m = 1/b_2$ ($-1 < m < 0$). Плотность вероятности этого типа распределений определяется формулой

$$f(x) = \begin{cases} \frac{m+1}{\alpha^{m+1}} (x + \alpha)^m, & \text{если } x \in [-\alpha, 0], \\ 0, & \text{если } x \notin [-\alpha, 0]. \end{cases}$$

Тип IX. $D < 0$, $\lambda < 0$ и $b_0 + 2b_1x + b_2x^2 = b_2(x + \alpha)x$, $\alpha = 2b_1/b_2$. Обозначим как $m = 1/b_2$ ($m < -1$). Плотность вероятности этого типа распределений определяется формулой

$$f(x) = \begin{cases} \frac{m+1}{\alpha^{m+1}} (x + \alpha)^m, & \text{если } x \in [-\alpha, 0], \\ 0, & \text{если } x \notin [-\alpha, 0]. \end{cases}$$

Тип X. $D = 0$, $\lambda = 0$ и $b_0 + 2b_1x + b_2x^2 = b_0$, числитель дроби в дифференциальном уравнении Пирсона равен a . Обозначим как $m = a/b_0 > 0$. Плотность вероятности этого типа распределений определяется формулой

$$f(x) = \begin{cases} me^{-mx}, & \text{если } x > 0, \\ 0, & \text{если } x \leq 0. \end{cases}$$

Это распределение является показательным.

Тип XI. $D = 0$, λ не определено, $b_0 + 2b_1x + b_2x^2 = b_0$. Обозначим как $\sigma^2 = b_0$. Плотность вероятности этого типа распределений определяется формулой

$$f(x) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{x^2}{2\sigma^2}}, \quad x \in (-\infty, \infty).$$

Данное распределение является нормальным.

Тип XII. Распределение этого типа совпадает с распределением типа I, если в последнем распределении положить $m = -n$.

Распределения Пирсона используются для сглаживания распределений выборочных значений. Эта система распределений применяется также для подбора распределения к эмпирическим данным тогда, когда по каким-либо причинам трудно или невозможно обосновать тип распределения генеральной совокупности. В этом случае по выборке вычисляются первые четыре момента, затем определяется тип распределения Пирсона, затем можно проверить степень согласованности эмпирических данных и полученного распределения Пирсона с помощью критерия Колмогорова-Смирнова или критерия χ^2 .

Основные статистические методы

При выполнении статистического анализа наиболее часто в качестве исходного материала используется последовательность независимых наблюдений случайной величины X . Другими словами, предполагается, что имеется вероятностный эксперимент, в котором наблюдается случайная величина X , и выполняется n независимых реализаций этого эксперимента. Наблюдаемые значения x_1, x_2, \dots, x_n называются *случайной выборкой*, количество наблюдений n — *объемом выборки*. Как случайная величина X , так и ее значения могут быть векторами. Множество возможных значений, которые могут наблюдаться при реализации эксперимента, образуют *выборочное пространство*, или, в других терминах, *генеральную совокупность*. С точки зрения теории вероятностей выборка x_1, x_2, \dots, x_n является реализацией некоторой случайной величины X . Задачи математической статистики возникают, когда функция распределения случайной величины X неизвестна, при этом методы статистического анализа позволяют получить информацию о различных закономерностях в генеральной совокупности.

Прежде чем описывать задачи статистического анализа, отметим, что перед непосредственным проведением анализа данных, как правило, выполняется этап предварительного анализа и обработки статистических данных. На этом этапе необходимо четко определить цели анализа, получить и первично обработать данные, определить их тип и структуру, подобрать и обосновать статистические методы, с помощью которых можно достичь целей анализа, подготовить данные для применения выбранных статистических методов и только после этого выполнить непосредственно статистический анализ данных. Этот этап, кроме формальных методов анализа данных, часто включает в себя неформальные способы оценки этих данных. Из сказанного ясно, что предварительный этап статистического анализа требует отдельного рассмотрения. Кроме того, на этом этапе также применяются статистические методы. Предварительному анализу посвящена глава 8. Здесь же мы рассмотрим общие понятия и методы математической статистики.

В зависимости от того, каков класс возможных распределений генеральной совокупности и что нужно знать о функции распределения, возникают различные статистические задачи. Рассмотрим основные из них.

2.1. Точечное оценивание параметров распределения

Предполагается, что неизвестная функция распределения принадлежит некоторому семейству распределений $F(u, \theta)$, зависящему от некоторого параметра θ (параметр θ , возможно, векторный, т.е. $\theta = (\theta_1, \theta_2, \dots, \theta_k)$); так, например,

семейство нормальных распределений зависит от двух параметров — математического ожидания и дисперсии. Нужно по наблюдениям (значениям выборки) оценить параметр (или несколько параметров).

Для построения оценок используются *статистики* — функции от выборочных значений. Распространенными примерами статистик являются:

$$\text{выборочное среднее } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i,$$

$$\text{выборочная дисперсия } S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2,$$

$$\text{выборочный } k\text{-й начальный момент } \bar{m}_k = \frac{1}{n} \sum_{i=1}^n x_i^k,$$

$$\text{выборочный } k\text{-й центральный момент } \bar{\mu}_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k.$$

Ниже будут приведены примеры других статистик.

Ясно, что не всякая статистика может служить оценкой неизвестного параметра распределения. Поскольку результаты опытов случайны, *любая статистика* представляет собой *случайную величину*. Чтобы статистика могла служить оценкой данного параметра θ , необходимо, чтобы распределение этой статистики было сосредоточено в достаточной близости от неизвестного значения параметра θ , т.е. так, чтобы вероятность больших отклонений этой статистики от θ была достаточно мала. Желательно также, чтобы точность оценивания увеличивалась при увеличении объема выборки. В связи с этим вводят следующие определения, характеризующие оценки.

Пусть $\hat{\theta}_n$ — некоторая статистическая оценка, полученная по выборке x_1, x_2, \dots, x_n и оценивающая неизвестный параметр θ распределения генеральной совокупности. Если оценка определяется одним числом $\hat{\theta}_n$, то ее называют *точечной*; если вычисляются две величины, θ_{1n} и θ_{2n} , такие, что $\theta_{1n} \leq \theta \leq \theta_{2n}$, то такую оценку для θ называют *интервальной* (интервальные оценки рассмотрены в следующем разделе).

2.1.1. Несмещенность оценки

Оценка $\hat{\theta}_n$ называется *несмещенной*, если при любом объеме выборки ее математическое ожидание равно оцениваемому параметру θ : $M\hat{\theta}_n = \theta$. Это свойство означает, что оценка $\hat{\theta}_n$ в среднем правильно оценивает неизвестный параметр θ ; т.е. если есть некоторое множество оценок данного параметра (значения одной и той же статистики), то среднее этих оценок будет совпадать с истинным значением параметра или будет к нему близко.

Отметим, что все выборочные начальные моменты, включая выборочное среднее, являются несмещенными оценками соответствующих моментов распределения генеральной совокупности. Однако выборочная дисперсия S_n^2 является

смещенной (как и другие центральные выборочные моменты): нетрудно показать, что $MS_n^2 = \frac{n-1}{n} DX$. Но поскольку при $n \rightarrow \infty$ $MS_n^2 \rightarrow DX$, оценки S_n^2 называют *асимптотически несмещенной*. Если немного изменить статистику S_n^2 , то новая оценка дисперсии будет несмещенной:

$$s_n^2 = \frac{n}{n-1} S_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Отметим также, что для несмещенных оценок величина $M(\hat{\theta}_n - \theta)^2$ совпадает с дисперсией статистики $\hat{\theta}_n$.

2.1.2. Эффективность оценки

Оценка $\hat{\theta}_n$ называется *эффективной*, если имеет наименьшую дисперсию среди всех возможных оценок параметра θ при фиксированном объеме выборки n . Эффективность оценки обеспечивает наименьший разброс возможных значений оценки $\hat{\theta}_n$ вокруг истинного значения параметра θ .

Эффективность оценок сильно зависит от распределения генеральной совокупности¹. Так, если генеральная совокупность имеет нормальное распределение, то выборочные среднее и дисперсия будут эффективными оценками.

2.1.3. Состоятельность оценки

Оценка $\hat{\theta}_n$ называется *состоятельной*, если при неограниченном росте объема выборки для произвольного $\varepsilon > 0$

$$P(|\hat{\theta}_n - \theta| > \varepsilon) \rightarrow 0 \text{ при } n \rightarrow \infty,$$

или, как говорят, $\hat{\theta}_n$ стремится к θ по вероятности при $n \rightarrow \infty$.

Понятия состоятельности оценки и несмещенности (точнее, асимптотической несмещенности) тесно связаны: если оценка $\hat{\theta}_n$ является состоятельной, то она асимптотически несмещенная. Обратное утверждение не верно, т.е. свойство состоятельности является более сильным, чем условие несмещенности.

Отметим, что выборочное среднее и выборочная дисперсия являются состоятельными оценками. Ниже будут приведены примеры выборочных статистик, которые часто используются в статистическом анализе, вместе с интервальными оценками параметров распределений, которые сейчас рассмотрим.

¹Точнее, эффективность оценок обычно доказывается (или, иначе говоря, строятся эффективные оценки) на основе метода максимального правдоподобия, в котором функция правдоподобия определяется исходя из предположения, что известен класс распределений, которому принадлежит распределение данной генеральной совокупности.

2.2. Интервальное оценивание параметров распределения

Точечные оценки имеют тот недостаток, что по ним нельзя судить о точности полученных оценок. Поэтому возникает задача определения на основании выборочных значений такого интервала (θ_1, θ_2) , который покрывал бы неизвестное значение параметра θ с заданной вероятностью.

Пусть $P(\theta_1 \leq \theta \leq \theta_2) = \alpha$, где случайный интервал (θ_1, θ_2) , который называется *доверительным интервалом*, с заданной вероятностью α содержит оцениваемый параметр θ . Величину α называют *доверительным уровнем* или *надежностью*. Величина $\delta = (\theta_1 - \theta_2)/2$ характеризует *точность* интервальной оценки. Обычно величину α берут равной 0,95, 0,99 или 0,999. Величину $1 - \alpha$ называют *уровнем значимости* отклонения оценки. Концы доверительного интервала θ_1 и θ_2 называют *доверительными границами*.

Один из распространенных методов построения доверительных интервалов заключается в следующем. По выборочным значениям вычисляется несмещенная точечная оценка $\hat{\theta}_n$ параметра θ . Напомним, что оценка (статистика) $\hat{\theta}_n$ является случайной величиной. Каким-либо способом вычисляется дисперсия статистики $\hat{\theta}_n$ или ее оценка $\hat{\sigma}_n^2$. Затем строится доверительный интервал вида $(\hat{\theta}_n - k_1 \hat{\sigma}_n, \hat{\theta}_n + k_2 \hat{\sigma}_n)$, где k_1 и k_2 — коэффициенты, значения которых определяют выбранный доверительный уровень и априорные предположения о распределении генеральной совокупности (например, нормальность или симметричность распределения). Но поскольку такой интервал определяется не однозначно, накладывается дополнительное условие, чтобы данный интервал имел минимальную длину. Если распределение статистики $\hat{\theta}_n$ симметрично (или близко к симметричному), то в этом случае доверительный интервал минимальной длины получается при $k_1 = k_2$. На такой основе строится, в частности, известный критерий Стьюдента (см. ниже) для нормально распределенных генеральных совокупностей. В самом общем случае (при минимальных предположениях относительно распределения генеральной совокупности) доверительные интервалы можно построить на основании неравенства Чебышева или других подобных неравенств (см. раздел 1.2.4). Однако такие интервальные оценки имеют небольшую точность.

Важную роль при построении точечных и интервальных оценок играют их асимптотические свойства. Часто явно или не явно используется следующая достаточно общая схема рассуждений [6, с. 371]. Пусть имеются независимые одинаково распределенные выборочные значения x_1, x_2, \dots, x_n , которые являются реализацией случайной величины X , имеющей функцию распределения $F(u)$. Требуется по выборке оценить математическое ожидание $G = MY = \int g(u) dF(u)$ случайной величины $Y = g(X)$. (Естественно предположение о том, что функция g такова, что случайная величина Y имеет конечный первый момент.) Статистика \hat{G}_n , оценивающая значение

величины G , вычисляется по формуле $\hat{G}_n = \frac{1}{n} \sum_{i=1}^n g(x_i)$. Эта оценка несмещена:

$$M\hat{G}_n = \frac{1}{n} \sum_{i=1}^n M g(x_i) = \frac{1}{n} \sum_{i=1}^n M g(X) = \frac{1}{n} \sum_{i=1}^n M Y = G.$$

По усиленному закону больших чисел она также состоятельна², т.е. с вероятностью 1 последовательность случайных величин \hat{G}_n сходится к значению G . Если еще потребовать, чтобы существовала дисперсия $DY = \sigma^2$, то из центральной предельной теоремы для суммы одинаково распределенных случайных слагаемых следует, что случайная величина

$$Z_n = \frac{\sqrt{n}}{\sigma} (\hat{G}_n - G) = \frac{1}{\sqrt{n}} \sum_{i=1}^n \frac{g(x_i) - G}{\sigma}$$

асимптотически нормальна с параметрами $(0, 1)$. Отсюда вытекает, что при больших n неравенство

$$\hat{G}_n - \alpha \frac{\sigma}{\sqrt{n}} < G < \hat{G}_n + \alpha \frac{\sigma}{\sqrt{n}}$$

выполняется с вероятностью $p_\alpha = \frac{2}{\sqrt{2\pi}} \int_0^\alpha e^{-\frac{u^2}{2}} du = 2\Phi(\alpha) - 1$, где $\Phi(u)$ — функция распределения стандартного нормального закона. При заданном значении вероятности p_α из последнего равенства определяется значение α , в результате получаем доверительный интервал

$$\left(\hat{G}_n - \alpha \frac{\sigma}{\sqrt{n}}, \hat{G}_n + \alpha \frac{\sigma}{\sqrt{n}} \right),$$

который содержит оцениваемое значение G с вероятностью p_α .

Построенная интервальная оценка не всегда применима на практике, поскольку значение дисперсии σ^2 может быть неизвестным. Однако при больших n , исходя из тех же соображений, которые изложены выше, имеем

$$DY = \sigma^2 \approx \frac{1}{n} \sum_{i=1}^n (g(x_i) - \hat{G}_n)^2 = S_n^2.$$

Ошибка, возникающая при замене в приведенных выше формулах величины σ ее оценкой S_n , имеет более высокий порядок малости (при $n \rightarrow \infty$), чем ошибка, возникающая при замене точного распределения случайной величины Z_n нормальным распределением. Поэтому “без зазрения совести” в качестве доверительного интервала для истинного значения G можно использовать интервал

$$\left(\hat{G}_n - \alpha \frac{S_n}{\sqrt{n}}, \hat{G}_n + \alpha \frac{S_n}{\sqrt{n}} \right).$$

Если одновременно оцениваются несколько параметров генеральной совокупности, то иногда возможно построение многомерных (размерность по числу оцениваемых параметров) доверительных областей, которые содержат неизвестные значения параметров. Однако построение таких областей вызывает определенные затруднения, поскольку статистики, оценивающие параметры, не являются независимыми случайными величинами (поэтому нельзя построить доверительную область просто как пересечение доверительных интервалов для отдельных параметров).

² Точнее, она сильно состоятельна.

Вместе с тем, если удастся построить такую доверительную область, как правило, она значительно более точно локализует значения неизвестных параметров распределения, чем простое пересечение доверительных интервалов. В этой книге мы не будем рассматривать многомерные доверительные области.

2.3. Выборочные статистики и интервальные оценки

Приведем примеры статистик и доверительные интервалы для них, которые находят наибольшее применение в статистическом анализе. В последующих частях книги будет показана их практическая реализация с использованием статистических функций или средств Excel.

Поскольку свойства этих статистик, и особенно способы построения доверительных интервалов, зависят от распределения генеральной совокупности, при их описании необходимо указывать *статистическую модель*, в рамках которой применимы данные статистики и доверительные интервалы. Статистическая модель описывает априорные предположения о распределении генеральной совокупности, требования к выборочным значениям (например, их независимость или минимальный объем выборки) и, возможно, способ представления данных. Статистические модели могут быть различными для точечных и интервальных оценок. Далее в этом разделе предполагается, что, если не указано другое, все выборочные значения независимы и имеют одинаковое распределение, т.е. требование независимости выборочных значений включается в статистическую модель обязательно.

Если конкретная выборка не соответствует определенной статистической модели, но все равно на основе данной выборки вычисляются какие-либо оценки, определяемые в рамках только этой статистической модели, то весьма вероятно, что те выводы, которые можно сделать на основе полученных оценок, окажутся ошибочными. Заметим, что статистический анализ выполняется не просто из-за любви к вычислениям, а для определенных целей, средством достижения которых служит статистический анализ. Причиной большинства неверных статистических выводов, которые весьма часто можно встретить на практике, является неправомерное применение оценок (и статистических критериев, о которых сказано ниже) в ситуации, когда выборка не удовлетворяет условиям статистической модели.

С другой стороны, часто можно пренебречь *умеренными* отклонениями от условий статистической модели и попытаться применить оценки данной модели. Поэтому при описании оценок будем показывать возможность ослабления условий статистической модели и указывать степень отклонения от модели, **при** которой статистики сохраняют свои свойства.

Сделаем еще одно замечание. Здесь мы не рассматриваем методы построения оценок и доверительных интервалов. Мы приводим только готовые формулы и рекомендации по их применению. Читатель, который хочет поближе познакомиться с методами построения оценок и доверительных интервалов, может обратиться к многочисленным изданиям по данному вопросу, среди которых выделим [5, 6, 17].

2.3.1. Статистика для оценивания математического ожидания

Точечная оценка

Статистическая модель. Приведенная ниже статистика применима для любого распределения генеральной совокупности, имеющего конечное математическое ожидание. Формула для дисперсии статистики правомерна при наличии

конечного второго момента у распределения генеральной совокупности. Здесь и далее, если не указано другое, n — объем выборки.

Статистика для оценки математического ожидания:

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Эта оценка несмещенная, эффективная и состоятельная. Ее дисперсия: $D\bar{x} = \frac{DX}{n}$,

где DX — дисперсия распределения генеральной совокупности. Отметим также, что коэффициент асимметрии $\beta_1(\bar{x})$ распределения статистики \bar{x} связан с коэффициентом асимметрии распределения генеральной совокупности $\beta_1(X)$ (см. раздел 1.2.3) зависимостью $\beta_1(\bar{x}) = \frac{\beta_1(X)}{\sqrt{n}}$.

дел 1.2.3) зависимостью $\beta_1(\bar{x}) = \frac{\beta_1(X)}{\sqrt{n}}$.

Интервальные оценки

Статистическая модель 1. Произвольное распределение генеральной совокупности с конечной известной дисперсией σ^2 .

Это наиболее общая статистическая модель. В рамках такой модели доверительный интервал для неизвестного математического ожидания можно построить только на основании неравенства Чебышева (см. раздел 1.2.4), которое в данном

случае будет иметь вид $P(|\bar{x} - MX| \leq k \frac{\sigma}{\sqrt{n}}) \leq 1 - \frac{1}{k^2}$. Коэффициент k находится

в соответствии с заданным доверительным уровнем α из равенства $\alpha = 1 - 1/k^2$:

$k = 1/\sqrt{1-\alpha}$. Доверительный интервал будет иметь вид $\left(\bar{x} - k \frac{\sigma}{\sqrt{n}}, \bar{x} + k \frac{\sigma}{\sqrt{n}} \right)$.

Статистическая модель 2. Генеральная совокупность имеет симметричное одномодальное распределение с известной конечной дисперсией σ^2 .

В этой статистической модели распределение статистики \bar{x} также будет симметричным и одномодальным³. Поэтому для построения интервальных оценок можно воспользоваться неравенством Гаусса, которое в данном случае будет

иметь вид $P(|\bar{x} - MX| \leq k \frac{\sigma}{\sqrt{n}}) \leq 1 - \frac{4}{9k^2}$. Значение k здесь вычисляется по формуле

$k = \frac{3}{2\sqrt{1-\alpha}}$, где α — заданный доверительный уровень, а доверительный интер-

вал имеет вид $\left(\bar{x} - k \frac{\sigma}{\sqrt{n}}, \bar{x} + k \frac{\sigma}{\sqrt{n}} \right)$.

Статистическая модель 3. Произвольное распределение генеральной совокупности с конечным четвертым моментом и неизвестной дисперсией. Объем выборки n больше 30.

³ К сожалению, в этой модели нельзя освободиться от условия симметричности распределения, поскольку, как известно, свертка одномодальных распределений не обязана быть одномодальным распределением. Другими словами, если распределение генеральной совокупности одномодально, то распределение \bar{x} не всегда будет одномодальным.

В данной модели можно построить интервальные оценки, основываясь на асимптотических свойствах статистики \bar{x} (см. раздел 2.2). Сделаем общее замечание о том, какой объем выборки считать достаточным, чтобы применять асимптотические оценки. В литературе по прикладной статистике обычно указывается, что для этого достаточно, чтобы n было больше 20, 25 или 30. Точную нижнюю границу для n определить сложно, поскольку она зависит от многих факторов, прежде всего от типа распределения. В теории вероятностей показано (неравенство Берри-Эссеена и подобные), что скорость сходимости распределения статистики \bar{x} к нормальному в равномерной метрике (и даже в интегральных метриках) имеет порядок $O\left(\frac{1}{\sqrt{n}}\right)$ и этот порядок нельзя улучшить, не вводя дополнительных предположений.

Отсюда следует, что значение n должно быть достаточно большим (хотя бы больше 100). Однако на практике уже при $n \geq 20$ получаются достаточно точные оценки. В дальнейшем, если не оговорено другое, будем применять асимптотические методы, когда $n \geq 30$. Однако следует помнить, что какой бы не был объем выборки, асимптотические оценки — это всегда только приближенные оценки.

Доверительный интервал в данной статистической модели строится на основе асимптотической нормальности оценок, как показано в разделе 2.2. Доверитель-

ный интервал имеет вид $\left(\bar{x} - k \frac{S_n}{\sqrt{n}}, \bar{x} + k \frac{S_n}{\sqrt{n}}\right)$, где коэффициент k определяется из

уравнения $\alpha = 2\Phi(k) - 1$, α — заданный доверительный уровень, Φ — функция распределения стандартного нормального закона.

Отметим, что применение вместо нормального распределения распределения Стьюдента расширяет доверительный интервал, тем самым повышая его надежность. Поэтому на практике обычно применяют доверительный интервал, построенный с помощью распределения Стьюдента, как показано в разделе 2.3.6.

Другие интервальные оценки для математического ожидания конкретных распределений будут показаны ниже.

2.3.2. Статистика для оценивания дисперсии

Точечная оценка

Статистическая модель. Приведенная ниже статистика применима для любого распределения генеральной совокупности, имеющего конечную дисперсию. Формула для дисперсии статистики правомерна при наличии конечного четвертого момента у распределения генеральной совокупности.

Статистика для оценки дисперсии DX:

$$\text{выборочная дисперсия } S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2,$$

где n — объем выборки.

Эта оценка асимптотически несмещенная, эффективная и состоятельная. Ее математическое ожидание равно $MS_n^2 = \frac{n-1}{n} DX$, дисперсия вычисляется по формуле

$$DS_n^2 = \frac{\mu_4 - \mu_2^2}{n} - \frac{2(\mu_4 - 2\mu_2^2)}{n^2} + \frac{\mu_4 - 3\mu_2^2}{n^3} = \frac{\mu_4 - \mu_2^2}{n} + O\left(\frac{1}{n^2}\right),$$

где μ_r — r -й центральный момент генеральной совокупности. Приведем также формулы для третьего центрального момента статистики S_n^2 и ее коэффициента асимметрии $\beta_1(S_n^2)$ (см. раздел 1.2.3):

$$M(S_n^2 - MS_n^2)^3 = \frac{\mu_6 - 3\mu_2\mu_4 - 6\mu_3^2 + 2\mu_2^3}{n^2} + O\left(\frac{1}{n^3}\right),$$

$$\beta_1(S_n^2) = \frac{\mu_6 - 3\mu_2\mu_4 - 6\mu_3^2 + 2\mu_2^3}{\sqrt{(\mu_4 - \mu_2^2)n}} + O\left(\frac{1}{n^{3/2}}\right).$$

Несмещенной оценкой для дисперсии DX будет статистика $s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$, дисперсия которой имеет порядок $DS_n^2 = \frac{\mu_4 - \mu_2^2}{n} + O\left(\frac{1}{n^2}\right)$.

Различие между оценками S_n^2 и s_n^2 имеет значение только при очень малых значениях n . При $n > 10$ разность между ними меньше 10%.

Приведем еще статистические характеристики оценки среднеквадратичного отклонения S_n :

$$MS_n = \sigma + O\left(\frac{1}{n}\right) \text{ и } DS_n = \frac{\mu_4 - \mu_2^2}{4\mu_2 n} + O\left(\frac{1}{n^2}\right).$$

Интервальные оценки

Статистическая модель. Произвольное распределение генеральной совокупности с конечным четвертым моментом. Объем выборки — не менее 50.

Если нет априорных предположений о типе распределения генеральной совокупности, то единственным способом построить доверительный интервал для неизвестной дисперсии является использование асимптотической нормальности распределения статистик для вычисления моментов генеральной совокупности. В этом случае доверительный интервал имеет вид $(S_n^2 - k\sigma(S_n^2), S_n^2 + k\sigma(S_n^2))$, где коэффициент k определяется из уравнения $\alpha = 2\Phi(k) - 1$, α — заданный доверительный уровень, Φ — функция распределения стандартного нормального закона. Среднеквадратическое отклонение $\sigma(S_n^2)$ статистики S_n^2 вычисляется по формуле $\sigma(S_n^2) = \sqrt{\frac{\bar{\mu}_4 - S_2^2}{n}}$, где $\bar{\mu}_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4$.

Другие интервальные оценки для дисперсии конкретных распределений будут показаны ниже.

⁴ По этой формуле вычисляется среднеквадратическое отклонение статистики S^* с точностью $O(\sqrt{4n})$. Можно использовать более точную формулу, но, как правило, этого не требуется.

2.3.3. Статистики для оценивания моментов

Точечные оценки для начальных моментов

Статистическая модель. Произвольное распределение генеральной совокупности с конечными моментами соответствующего порядка.

Статистика для оценки начального момента m_k порядка k :

$$\text{выборочный } k\text{-й начальный момент } \bar{m}_k = \frac{1}{n} \sum_{i=1}^n x_i^k.$$

Эта оценка несмещенная, эффективная и состоятельная. Распределение статистики \bar{m}_k асимптотически нормально. Ее дисперсия: $D\bar{m}_k = \frac{m_{2k} - m_k^2}{n}$.

Точечные оценки для центральных моментов

Статистическая модель. Произвольное распределение генеральной совокупности с конечными моментами соответствующего порядка.

Статистика для оценки центрального момента μ_k порядка k :

$$\text{выборочный } k\text{-й центральный момент } \bar{\mu}_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k.$$

Эта оценка асимптотически несмещенная: $M\bar{\mu}_k = \mu_k + O(n^{-1})$. Ее дисперсия:

$$D\bar{\mu}_k = \frac{\mu_{2k} - 2k\mu_{k-1}\mu_{k+1} - \mu_k^2 + k^2\mu_2\mu_{k-2}}{n} + O\left(\frac{1}{n^2}\right).$$

Интервальные оценки для моментов выше второго порядка строятся редко. Если не делать предположений о типе распределения генеральной совокупности, то доверительные интервалы для неизвестных моментов можно построить только на основе их асимптотической нормальности при достаточно больших значениях n .

2.3.4. Статистики для оценивания коэффициентов асимметрии и эксцесса

Точечные оценки

Статистическая модель. Произвольное распределение генеральной совокупности с конечными моментами четвертого порядка.

Напомним, что коэффициент асимметрии вычисляется по формуле $\beta_1 = \mu_3/\mu_2^{3/2}$, а коэффициент эксцесса — по формуле $\beta_2 = \mu_4/\mu_2^2 - 3$, где μ_k — центральные моменты порядка k (см. раздел 1.2.3). Для получения оценок этих коэффициентов

вычисляются выборочные центральные моменты $\bar{\mu}_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$, которые затем подставляются в приведенные формулы вместо μ_k . Получаем оценки:

$$\bar{\beta}_1 = \frac{\bar{\mu}_3}{\sqrt{\bar{\mu}_2^3}}, \quad \bar{\beta}_2 = \frac{\bar{\mu}_4}{\bar{\mu}_2^2} - 3.$$

Эти оценки состоятельные и асимптотически несмещенные: $M\bar{\beta}_1 = \beta_1 + O(n^{-1})$, $M\bar{\beta}_2 = \beta_2 + O(n^{-1})$. Отметим также, что $D\bar{\beta}_1 = \frac{d}{n} + O\left(\frac{1}{n^{3/2}}\right)$, где

$$d = \frac{4\mu_2^2\mu_6 - 12\mu_2\mu_3\mu_5 - 24\mu_2^3\mu_4 + 9\mu_3^2\mu_4 + 35\mu_2^2\mu_3^2 + 36\mu_2^5}{4\mu_2^5}.$$

Если распределение генеральной совокупности симметрично, то

$$d = \frac{4\mu_2^2\mu_6 - 24\mu_2^3\mu_4 + 36\mu_2^5}{4\mu_2^5}.$$

Отметим также, что на практике, если выборочное распределение близко к нормальному, для оценки среднеквадратических отклонений s_1 и s_2 коэффициентов $\bar{\beta}_1$ и $\bar{\beta}_2$ используют формулы

$$s_1 = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}}, \quad s_2 = \sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}.$$

В разделе 9.2 показано практическое применение коэффициентов $\bar{\beta}_1$ и $\bar{\beta}_2$ для подбора функций распределений.

2.3.5. Статистика для оценивания медианы

Точечная оценка

Статистическая модель. Произвольное распределение генеральной совокупности.

Напомним, что медианой называют такое значение m , которое делит распределение на две равновероятные половины, т.е. $P(X < m) = P(X \geq m) = 1/2$.

Точечная оценка для медианы строится следующим образом.

На основании выборочных значений строится вариационный ряд, т.е. значения x_1, x_2, \dots, x_n располагаются в порядке возрастания. Получаем последовательность $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$, которая называется *вариационным рядом* (о вариационном ряде речь идет ниже, в разделе 2.3.9). Если объем выборки n — нечетное число, т.е. $n = 2k + 1$, то в качестве оценки медианы выбирается значение $x_{(k)}$ из вариационного ряда. Если n четное, т.е. $n = 2k$, то в качестве оценки медианы выбирается полусумма значений $x_{(k)}$ и $x_{(k+1)}$ вариационного ряда⁵. Более подробно получение оценки медианы описано в разделе 8.4.

Далее приведем методы оценивания параметров некоторых конкретных распределений, для реализации которых в Excel предусмотрены специальные функции или средства (см. главы 4 и 5).

2.3.6. Оценки параметров нормального распределения

Статистическая модель. Генеральная совокупность имеет нормальное распределение с математическим ожиданием m и дисперсией σ^2 .

⁵ В принципе, в последнем случае в качестве оценки медианы можно взять любое значение из интервала $(X_{(k)}, X_{(k+1)})$.

Точечные оценки

Для точечного оценивания математического ожидания m и дисперсии σ^2 используются статистики $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ и $S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$, которые являются несмещенными (S_n^2 асимптотически несмещенная), эффективными и состоятельными оценками неизвестных m и σ^2 . Формулы для дисперсий этих статистик приведены в разделах 2.3.1 и 2.3.2.

Статистика \bar{x} распределена по нормальному закону с математическим ожиданием m и дисперсией σ^2/n , а случайная величина $\sqrt{n}(\bar{x} - m)/\sigma$ имеет стандартное нормальное распределение. Случайная величина $n S_n^2/\sigma^2$ имеет распределение χ^2 с $(n-1)$ степенью свободы (см. раздел 1.5.5). Распределение Стьюдента с $(n-1)$ степенью свободы (см. раздел 1.5.6) имеет случайная величина $\sqrt{n-1}(\bar{x} - m)/S_n$. Эти свойства статистик \bar{x} и S_n^2 используются для построения доверительных интервалов.

Интервальные оценки для математического ожидания

Способ построения доверительного интервала для математического ожидания зависит от того, известно ли значение дисперсии σ^2 . Если значение дисперсии известно, то доверительный интервал, соответствующий доверительному уровню

α , имеет вид $\left(\bar{x} - k \frac{\sigma}{\sqrt{n}}, \bar{x} + k \frac{\sigma}{\sqrt{n}} \right)$, где коэффициент k определяется из уравнения

$\alpha = 2\Phi(k) - 1$, Φ — функция распределения стандартного нормального закона.

(В Excel, кроме других средств, для построения доверительного интервала можно воспользоваться функцией Excel ДОВЕРИТ, которая по заданным значениям

α , σ и n вычисляет величину $k \frac{\sigma}{\sqrt{n}}$ (см. раздел 4.11.2).)

В случае, когда значение дисперсии σ^2 неизвестно, вместо этого значения используют выборочную дисперсию $S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$, а значение коэффициента k

определяется из уравнения $\alpha = 2F_{n-1}(k) - 1$, где F_{n-1} — функция распределения Стьюдента с $(n-1)$ степенью свободы, поскольку случайная величина $\sqrt{n-1}(\bar{x} - m)/S_n$ имеет именно такое распределение. Доверительный интервал

имеет вид $\left(\bar{x} - k \frac{S_n}{\sqrt{n-1}}, \bar{x} + k \frac{S_n}{\sqrt{n-1}} \right)$.

В главе 10 показана практическая реализация построения доверительных интервалов для математического ожидания.

Интервальные оценки для дисперсии

Предположим, что математическое ожидание m и дисперсия σ^2 распределения генеральной совокупности неизвестны (случай, когда известно математическое ожидание, рассмотрен в главе 10). Поскольку случайная величина $n S_n^2/\sigma^2$ имеет

распределение χ^2 с $(n-1)$ степенью свободы, доверительный интервал для σ^2 при заданном доверительном уровне α строится следующим образом. Вычисляются точечные оценки $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ и $S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$ и определяются коэффициенты $t_n = F_{n-1}^{-1}(\beta_n)$ и $t_n = F_{n-1}^{-1}(\beta_n)$, где $\beta_n = (1 - \alpha)/2$, $\beta_n = (1 + \alpha)/2$, F_{n-1}^{-1} — функция, обратная к функции распределения χ^2 с $(n-1)$ степенью свободы. Доверительный интервал имеет вид $\left(\frac{n}{t_n} S_n^2, \frac{n}{t_n} S_n^2 \right)$.

Также можно построить *доверительную область* для совместного оценивания математического ожидания и дисперсии [7, с. 94, 18, с. 181].

2.3.7. Оценка параметра p распределения Бернулли

Напомним, что распределение Бернулли обычно рассматривается как модель случайного эксперимента, в результате которого с вероятностью p может произойти исход “1” и с вероятностью $1 - p$ — исход “0” (см. раздел 1.4.2). Целью статистического анализа обычно является определение значения вероятности p (вероятность p часто называют *биномиальной вероятностью*).

Статистическая модель 1. Выборка x_1, x_2, \dots, x_n является результатом наблюдения за одним экспериментом, состоящим из n одинаковых испытаний, в каждом из которых с вероятностью p может произойти исход “1” и с вероятностью $(1 - p)$ — исход “0”. Здесь x_i равно 1, если в i -м испытании произошел исход “1”, и 0 — в противном случае.

Точечная оценка

Несмещенной и эффективной оценкой для вероятности p будет статистика $\hat{p} = r/n$, где r — количество исходов “1”. Дисперсия статистики \hat{p} : $D\hat{p} = p(1 - p)/n$, ее выборочная оценка: $S_n^2(\hat{p}) = r(n - r)/n^2(n - 1)$. Случайная величина r имеет биномиальное распределение с параметрами n и p (см. раздел 1.4.3). Распределение статистики \hat{p} асимптотически нормально с параметрами $m = p$ и $\sigma^2 = p(1 - p)/n$.

Интервальные оценки

Доверительные интервалы для неизвестного значения вероятности p строятся или на основе биномиального распределения, которое имеет случайная величина r , или на основе асимптотической нормальности распределения статистики \hat{p} .

Доверительный интервал на основе биномиального распределения для значения вероятности p при заданном доверительном уровне α строится следующим образом⁶. Сначала подсчитывается величина r — количество исходов “1”, затем определяются коэффициенты $t_n = F_{k1,k2}^{-1}(\beta_n)$ и $t_n = F_{k3,k4}^{-1}(\beta_n)$, где $\beta_n = (1 - \alpha)/2$, $\beta_n = (1 + \alpha)/2$, $F_{m1,m2}^{-1}$ — функция, обратная к функции бета-распределения

с параметрами m_1 и m_2 (см. раздел 1.5.9), $k_1 = r$, $k_2 = n - r + 1$, $k_3 = r + 1$, $k_4 = n - r$. Доверительный интервал имеет вид (t_*, t_*) . Здесь использованы известные соотношения между биномиальным распределением и бета-распределением: если X — случайная величина, имеющая биномиальное распределение с параметрами n и p , тогда $P(X \leq k) = F_{n-k, k+1}(1-p)$, где $F_{n-k, k+1}$ — функция бета-распределения с соответствующими параметрами.

Доверительный интервал на основе асимптотической нормальности. Как указывалось, при достаточно большом n ($n \geq 30$) точечная оценка $\hat{p} = r/n$ распределена приближенно по нормальному закону с математическим ожиданием p и дисперсией $p(1-p)/n$. Поэтому приближенный доверительный интервал для значения вероятности p можно построить следующим образом. Вычисляется точечная оценка $\hat{p} = r/n$.

При заданном доверительном уровне α из уравнения $\alpha = 2\Phi(k) - 1$, где Φ — функция распределения стандартного нормального закона, определяется значение k . Далее можно построить доверительные интервалы двух типов: более точный интервал

$$\left(\frac{\hat{p}n + \frac{1}{2}k^2 - k\sqrt{\hat{p}(1-\hat{p})n + \frac{1}{4}k^2}}{n + k^2}, \frac{\hat{p}n + \frac{1}{2}k^2 + k\sqrt{\hat{p}(1-\hat{p})n + \frac{1}{4}k^2}}{n + k^2} \right)$$

и более простой, но менее точный, интервал вида

$$\left(\hat{p} - k\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + k\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right).$$

Здесь при построении первого доверительного интервала используется только аппроксимация биномиального распределения нормальным, при построении второго неизвестное значение дисперсии $D\hat{p} = p(1-p)/n$ заменяется величиной $\hat{p}(1-\hat{p})/n$.

Статистическая модель 2. Выборка x_1, x_2, \dots, x_n состоит из результатов n экспериментов, в каждом из которых проводилось N испытаний, в каждом испытании с вероятностью p может произойти исход “1” и с вероятностью $(1-p)$ — исход “0”. Здесь x_i равно числу исходов “1” в i -м эксперименте.

Точечная оценка

Несмещенной и эффективной оценкой для вероятности p будет статистика $\hat{p} = \frac{1}{nN} \sum_{i=1}^n x_i$. Дисперсия статистики \hat{p} : $D\hat{p} = p(1-p)/nN$. Распределение статистики \hat{p} асимптотически нормально с параметрами $m = p$ и $\sigma^2 = p(1-p)/nN$.

Интервальные оценки

Поскольку значение величины nN , как правило, больше 30, то наиболее простой доверительный интервал для неизвестного значения вероятности p строится на основе асимптотической нормальности распределения статистики \hat{p} , которая

здесь вычисляется по формуле $\hat{p} = \frac{1}{nN} \sum_{i=1}^n x_i$. По заданному значению доверительного

уровня α из уравнения $\alpha = 2\Phi(k) - 1$, где Φ — функция распределения стандартного нормального закона, определяется значение коэффициента k . Доверительный интервал имеет вид

$$\left(\hat{p} - k \sqrt{\frac{\hat{p}(1-\hat{p})}{nN}}, \hat{p} + k \sqrt{\frac{\hat{p}(1-\hat{p})}{nN}} \right).$$

Здесь, как и в асимптотических оценках предыдущей модели, при построении доверительного интервала используется аппроксимация биномиального распределения нормальным и замена неизвестного значения дисперсии $D\hat{p} = p(1-p)/nN$ величиной $\hat{p}(1-\hat{p})/nN$.

Преобразование арксинуса

Недостатком асимптотических доверительных интервалов является то, что при их построении неизвестное значение дисперсии $D\hat{p}$ заменяется величиной $\hat{p}(1-\hat{p})/n$ (в модели 1) или величиной $\hat{p}(1-\hat{p})/nN$ (в модели 2). Существует преобразование статистики \hat{p} , распределение которого почти не зависит от неизвестного значения вероятности p . Такое преобразование называется *преобразованием арксинуса* и имеет вид $z = \arcsin \sqrt{\hat{p}}$. Математическое ожидание случайной величины z приближенно равно $\arcsin \sqrt{p}$, а дисперсия приближенно равна $1/4n$. Кроме того, распределение величины z ближе к нормальному, чем распределение статистики \hat{p} .

Иногда используют другой вариант преобразования арксинуса: $y = 2\sqrt{n} \arcsin \sqrt{\hat{p}}$. Здесь дисперсия случайной величины y практически не зависит от n и p и приближенно равна 1. Ее математическое ожидание приближенно равно $2\sqrt{n} \arcsin \sqrt{p}$.

Приведенные преобразования не применимы в случае, когда значение p близко к 0 или 1. Преобразование Энскомба $w = \arcsin \sqrt{\frac{r+3/8}{n+1/4}}$ (r — количество исходов "1") лишено этого недостатка. Дисперсия случайной величины w приближенно равна $1/(4n+2)$.

Практическое построение доверительных интервалов на основе преобразования арксинуса показано в главе 10, раздел 10.8.3.

2.3.8. Оценка параметра λ распределения Пуассона

Статистическая модель. Генеральная совокупность имеет распределение Пуассона с параметром λ (см. раздел 1.4.4).

Точечная оценка

Выборочное среднее $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ будет несмещенной и эффективной оценкой для неизвестного параметра λ . Дисперсия этой оценки: $D\bar{x} = \lambda/n$. Случайная

величина $\sum_{i=1}^n x_i$ имеет распределение Пуассона с параметром $n\lambda$, а случайная величина $\sqrt{n/\lambda}(\bar{x} - \lambda)$ асимптотически нормальна с параметрами $(0, 1)$.

Интервальные оценки

Доверительные интервалы для неизвестного значения вероятности λ строятся или на основе распределения Пуассона, которое имеет случайная величина $\sum_{i=1}^n x_i$, или на основе асимптотической нормальности распределения случайной величины $\sqrt{n/\lambda}(\bar{x} - \lambda)$.

Использование распределения Пуассона. Если задан доверительный уровень α и вычислена точечная оценка $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, то далее для построения доверительного интервала с использованием распределения Пуассона определяются коэффициенты $t_n = F_k^{-1}(\beta_n)$ и $t_{\bar{n}} = F_k^{-1}(\beta_{\bar{n}})$, где $\beta_n = (1 - \alpha)/2$, $\beta_{\bar{n}} = (1 + \alpha)/2$, F_k^{-1} — функция, обратная к функции χ^2 -распределения с $k = 2(n\bar{x} + 1)$ степенями свободы (см. раздел 1.5.5). Доверительный интервал имеет вид $\left(\frac{t_n}{2n}, \frac{t_{\bar{n}}}{2n}\right)$.

Здесь использовано соотношение между распределением Пуассона и распределением χ^2 (см. раздел 1.4.4): $P(X \leq k) = P(Z \geq 2\lambda)$, где X — случайная величина, распределенная по закону Пуассона с параметром λ , Z — случайная величина, имеющая χ^2 -распределение с $2(k + 1)$ степенью свободы.

Использование асимптотической нормальности. При достаточно большом n приближенный доверительный интервал для значения λ строится таким образом. Задается доверительный уровень α и вычисляется точечная оценка $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$. Из уравнения $\alpha = 2\Phi(k) - 1$, где Φ — функция распределения стандартного нормального закона, определяется значение k . Можно построить доверительные интервалы двух типов: более точный интервал

$$\left(\bar{x} + \frac{k^2}{2n} - \frac{k}{2n} \sqrt{k^2 + 4n\bar{x}}, \bar{x} + \frac{k^2}{2n} + \frac{k}{2n} \sqrt{k^2 + 4n\bar{x}} \right)$$

и более простой, но менее точный, интервал вида

$$\left(\bar{x} - k \sqrt{\frac{\bar{x}}{n}}, \bar{x} + k \sqrt{\frac{\bar{x}}{n}} \right).$$

При построении первого доверительного интервала используется только аппроксимация распределения Пуассона нормальным, при построении второго неизвестное значение дисперсии $D\bar{x} = \lambda/n$ заменяется величиной \bar{x}/n .

2.3.9. Порядковые статистики

Порядковые (ранговые) статистики играют большую роль в математической статистике. На их основе строятся так называемые *непараметрические* или *свободные от распределения* методы, т.е. методы, которые не зависят от неизвестного распределения генеральной совокупности. Некоторые такие методы будут описаны ниже, в разделе 2.4. Кроме того, порядковые статистики используются для построения эмпирической функции распределения, аппроксимирующей распределение генеральной совокупности (см. раздел 8.3), для оценивания квантилей распределения генеральной совокупности, как показано ниже в этом разделе, и во многих других статистических методах.

Статистическая модель. Имеется конечная выборка x_1, x_2, \dots, x_n объемом n , которая является реализацией случайной величины X с функцией распределения $F(u)$.

Упорядоченная по возрастанию последовательность выборочных значений

$$x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$$

называется *вариационным рядом*. Равные между собой члены выборки нумеруются в произвольном порядке. Члены вариационного ряда $x_{(i)}$ ($i = 1, 2, \dots, n$) называются *порядковыми (ранговыми) статистиками*. Число $r_i = i$ называется *рангом* члена $x_{(i)}$. (В литературе также можно встретить определение ранга как $r_i = i/n$.) В статистическом анализе часто используется статистика $R_n = x_{(n)} - x_{(1)}$, которая называется *размахом* или *широтой* выборки.

Распределение порядковых статистик зависит от распределения генеральной совокупности, но вероятность того, что случайная величина X примет значение из интервала $(x_{(i-1)}, x_{(i)})$, не зависит от распределения и всегда равна $1/(n+1)$. В частности, вероятности $P(X < x_{(0)})$ и $P(X > x_{(n)})$ также равны $1/(n+1)$ [6, с. 367].

Оценки квантилей

Напомним, что квантилью порядка p случайной величины X называется такое число ξ_p , что $P(X < \xi_p) = p$. (Медиана является квантилью порядка 0,5).

Оценкой неизвестной квантили порядка p ξ_p принимается выборочная p -квантиль $\bar{\xi}_p = x_{(k(p))}$, где $k(p) = np$, если np — целое число и $k(p) = [np] + 1$ в противном случае⁷.

2.4. Проверка статистических гипотез

Статистической гипотезой называется утверждение, высказанное относительно распределения генеральной совокупности или некоторых его параметров. Обычно такую гипотезу обозначают как H_0 — это *нулевая* (предложенная) *гипотеза*. Противоположное утверждение — отрицание гипотезы H_0 — называется *конкурирующей* (или *альтернативной*) *гипотезой* и обозначается как H_1 .

Приведем несколько примеров статистических гипотез.

Гипотеза H_0 : выборка x_1, x_2, \dots, x_n получена из генеральной совокупности, равномерно распределенной на интервале $[a, b]$.

Гипотеза H_0 : выборочные значения извлечены из генеральной совокупности, математическое ожидание которой лежит в пределах от a до b (a и b — априорно заданные числа).

[7] обозначает целую часть числа t .

Имеем парные наблюдения $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, являющиеся реализацией случайной величины $Z = (X, Y)$. Гипотеза H_0 : компоненты X и Y независимы.

Есть две выборки, x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_k , извлеченные из двух генеральных совокупностей с неизвестными математическими ожиданиями m_1 и m_2 соответственно. Гипотеза H_0 : $m_1 \geq m_2$.

Очевидно, в каждой конкретной ситуации можно сформулировать целое семейство различных гипотез. При проведении статистического анализа из этого множества гипотез следует выбрать те гипотезы, которые сформулированы наиболее четко, не оставляя места двойственности в утверждениях, и максимально соответствуют цели конкретного исследования. Рекомендуется также выбирать простые гипотезы, сформулированные относительно одного параметра распределения, так как сложные гипотезы требуют и сложных критериев для проверки их истинности.

Критерий проверки статистической гипотезы — это процедура выработки решения о том, принять или отвергнуть данную гипотезу. *Критической областью* критерия (или *областью непринятия гипотезы*) является та часть выборочного пространства, которая приводит к отклонениям гипотезы. *Уровнем значимости α* критерия является вероятность того, что этот критерий приведет к отклонению нулевой гипотезы в случае ее истинности, т.е. вероятность того, что при выполнении нулевой гипотезы результаты проверок попадут в критическую область. Если результаты проверки находятся в критической области, то гипотеза H_0 отклоняется и принимается альтернативная гипотеза H_1 . Поэтому критическая область должна быть расположена там, где она соответствует конкурирующей гипотезе. При выборе гипотез нулевой гипотезой (по сравнению с альтернативной) должна быть та гипотеза, которую более опасно ошибочно отвергнуть.

Отклонение нулевой гипотезы в случае ее истинности называется *ошибкой первого рода*. Поэтому уровень значимости α есть вероятность совершения ошибки первого рода. Принятие гипотезы H_0 , когда она неверна, называется *ошибкой второго рода*. Вероятность ошибки второго рода обычно обозначают как β .

Естественно стремление минимизировать вероятности ошибок первого и второго рода. Снижая уровень значимости α , тем самым снижаем вероятность возникновения ошибки первого рода, но в этом случае возрастает вероятность β возникновения ошибок второго рода. В связи с этим вводят понятие *мощности критерия*, которое определяют как вероятность отклонения нулевой гипотезы, когда она неверна, т.е. мощность критерия можно определить как $1 - \beta$. Эта вероятность зависит от реального значения рассматриваемого параметра генеральной совокупности. Поскольку реальное значение параметра заранее не известно, рассматривают функцию мощности, которая показывает соответствующее значение мощности критерия для каждого возможного значения параметра. Функция мощности играет в теории проверки гипотез фундаментальную роль. Она полностью характеризует критерий, так как показывает, насколько хорошо он соответствует своему основному назначению — «улавливать» возможные отклонения от нулевой гипотезы.

Часто возможные значения критериальной статистики, на основе которой строится критерий, принадлежат некоторому интервалу. Тогда критическая область также является интервалом. Граничные точки критической области называются *критическими значениями*. Критические значения выбираются таким образом, чтобы при выбранном уровне значимости α мощность критерия $1 - \beta$ была наибольшей.

1. Правосторонняя критическая область в виде интервала $(t_{кр}, +\infty)$, где критическое значение $t_{кр}$ определяется из равенства $P(\theta > t_{кр}) = \alpha$. Значение $t_{кр}$ называется *правосторонней критической точкой*, отвечающей уровню значимости α .
2. Левосторонняя критическая область в виде интервала $(-\infty, t_{кр})$, где критическое значение $t_{кр}$ определяется из равенства $P(\theta < t_{кр}) = \alpha$. Значение $t_{кр}$ называется *левосторонней критической точкой*, отвечающей уровню значимости α .
3. Двухсторонняя критическая область, состоящая из двух интервалов $(-\infty, t_{кр1})$ и $(t_{кр2}, +\infty)$, где критические значения $t_{кр1}$ и $t_{кр2}$ определяется из равенств $P(\theta < t_{кр1}) = \alpha/2$ и $P(\theta > t_{кр2}) = \alpha/2$. Эти значения называются *двухсторонними критическими точками*, отвечающими уровню значимости α .

Возможны три варианта расположения критической области, определяемых видом нулевой и альтернативной гипотез, а также распределением критериальной статистики 0.

1. Правосторонняя критическая область в виде интервала $(t_{кр}, +\infty)$, где критическое значение $t_{кр}$ определяется из равенства $P(\theta > t_{кр}) = \alpha$. Значение $t_{кр}$ называется *правосторонней критической точкой*, отвечающей уровню значимости α .
2. Левосторонняя критическая область в виде интервала $(-\infty, t_{кр})$, где критическое значение $t_{кр}$ определяется из равенства $P(\theta < t_{кр}) = \alpha$. Значение $t_{кр}$ называется *левосторонней критической точкой*, отвечающей уровню значимости α .
3. Двухсторонняя критическая область, состоящая из двух интервалов $(-\infty, t_{кр1})$ и $(t_{кр2}, +\infty)$, где критические значения $t_{кр1}$ и $t_{кр2}$ определяется из равенств $P(\theta < t_{кр1}) = \alpha/2$ и $P(\theta > t_{кр2}) = \alpha/2$. Эти значения называются *двухсторонними критическими точками*, отвечающими уровню значимости α .

Необходимо подчеркнуть, что статистические критерии на основании выборочных наблюдений *не доказывают* ту или иную гипотезу. Они позволяют утверждать, что выборочные значения *не противоречат* принятой гипотезе. Таким образом, выводы, принимаемые на основе статистических данных, формулируются в следующем виде: "экспериментальные данные согласуются с данной гипотезой (или противоречат ей)".

Следует предупредить об опасности, связанной с применением нескольких статистических критериев при анализе одних и тех же данных. Если к одним и тем же данным применяют два различных критерия для проверки одной и той же нулевой гипотезы (или двух сходных гипотез) и в каждом случае принимается уровень значимости, например, 0,05, то вероятность того, что хотя бы по одному из критериев нулевая гипотеза будет ошибочно отклонена, превосходит 0,05. Следует воспользоваться лишь одним критерием, желательно более мощным.

Сделаем еще некоторые замечания о применении статистических критериев. Все статистические критерии для корректного своего использования предполагают выполнение некоторого комплекса условий (например, условия нормальности распределения генеральной совокупности). Глава 2. Основные статистические методы работы в рамках конкретной статистической модели. На практике условия, налагаемые статистической моделью, могут и не выполняться, что приводит к возрастанию вероятности неправильных выводов, которые делаются на основе того или иного кри-

о каком-либо конкретном виде распределения генеральной совокупности. При исследовании выборки из нормально распределенной генеральной совокупности критерии этого типа несколько уступают по мощности соответствующим критериям, построенным на предположении о нормальности. Они обладают, однако, тем преимуществом, что свободны от подобного предположения о нормальности, поэтому их можно использовать в ситуациях, когда вид распределения заранее не известен.

Чтобы показать, как строятся и как “работают” критерии проверки гипотез, рассмотрим три типа критериев: критерии проверки гипотез о значениях параметров генеральной совокупности, критерии проверки гипотез о различии (или равенстве) параметров нескольких генеральных совокупностей и критерии проверки гипотез о принадлежности распределения генеральной совокупности определенному классу распределений. Многочисленные примеры других критериев приведены в частях III и IV книги.

2.4.1. Критерии проверки гипотез о значениях параметров генеральной совокупности

Многие подобные критерии строятся на основе доверительных интервалов (см. раздел 2.3). Например, необходимо проверить гипотезу, что неизвестное математическое ожидание m генеральной совокупности равно некоторому конкретному значению m_0 . Пусть на основе выборки построен доверительный интервал (t_1, t_2) с доверительным уровнем α (т.е. с вероятностью α этот интервал содержит неизвестное значение математического ожидания m). Тогда, если интервал (t_1, t_2) покрывает значение m_0 (т.е. выполняются неравенства $t_1 \leq m_0 \leq t_2$), принимается выдвинутая гипотеза с уровнем значимости $1 - \alpha$. Критическая область здесь состоит из объединения двух областей: $(-\infty, t_1)$ и $(t_2, +\infty)$, t_1 и t_2 являются двухсторонними критическими точками. Таким образом, чтобы проверить гипотезу $H_0: m = m_0$ с уровнем значимости $1 - \alpha$, необходимо построить для значения m доверительный интервал (t_1, t_2) с доверительным уровнем α и проверить выполнение неравенств $t_1 \leq m_0 \leq t_2$. Если эти неравенства выполняются, то с вероятностью α гипотеза H_0 принимается. Если хотя бы одно из этих неравенств не выполняется, то гипотеза отклоняется. Аналогично строятся критерии о проверке гипотез в виде неравенств. Например, гипотеза $H_0: m \geq m_0$. В этом случае необходимо построить правосторонний доверительный интервал вида $(t, +\infty)$, который содержал бы значение m с вероятностью α , и проверить неравенство $t \geq m_0$ (t — критическое значение). Если это неравенство выполняется, то гипотеза H_0 принимается с уровнем значимости $1 - \alpha$. В противном случае она отвергается.

Обычно в подобных критериях для упрощения вычислений доверительные интервалы строятся не для неизвестного параметра распределения, а для унифицированной статистики, которая при условии истинности гипотезы H_0 имеет известное распределение. Такая статистика называется *критериальной статистикой*. Например, для критерия проверки гипотезы о значении математического

ожидания, сформулированной выше, вычисляется статистика
$$T = \frac{\sqrt{n}(\bar{x} - m_0)}{S_n}$$

(здесь $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ и $S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$), которая в случае нормального распределе-

ния генеральной совокупности и при выполнении условия $m = m_0$ подчиняется распределению Стюдента. Тогда границами критической области для критерия будут просто квантили этого распределения, порядок которых определяется заданным уровнем значимости. На таком принципе построены приведенные ниже критерии. Большую роль в таких критериях играют априорные предположения о распределении генеральной совокупности. Поэтому использование конкретных критериев проверки гипотез требует обязательного соблюдения условий статистической модели, в рамках которой применим данный критерий.

Во многих случаях, чтобы уменьшить зависимость критериальных статистик от априорных предположений о распределении генеральной совокупности (а также в случаях, когда с точными распределениями критериальных статистик по каким-либо причинам трудно работать (сложные вычисления и т.п.)), критерии строят на основе асимптотических распределений этих статистик. При использовании таких критериев следует помнить, что, во-первых, они работают только при достаточно большом объеме выборки, во-вторых, эти критерии приближенные, степень точности которых удается определить только в редких случаях.

Приведем несколько критериев проверки значений параметров распределения, которые строятся на основе доверительных интервалов. Рассмотрим критерии для случая равенства и неравенств. Формы описания критериев, используемой в этих примерах, будем придерживаться и далее при описании критериев в последующих главах.

Критерий проверки значения математического ожидания нормальной совокупности

Статистическая модель. Выборка x_1, x_2, \dots, x_n получена из генеральной совокупности, подчиняющейся нормальному закону распределения с неизвестным математическим ожиданием μ и неизвестной дисперсией σ^2 .

Гипотезы

а) Равенство

$$H_0: \mu = m_0$$

$$H_1: \mu \neq m_0$$

б) Неравенство

$$H_0: \mu < m_0$$

$$H_1: \mu > m_0$$

в) Неравенство

$$H_0: \mu > m_0$$

$$H_1: \mu < m_0$$

Здесь m_0 — заданное число. Задан уровень значимости α .

В качестве критериальной статистики используем статистику $T = \frac{\sqrt{n}(\bar{x} - m_0)}{s}$,

где $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, и $s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$. При условии истинности гипотезы H_0 статистика T имеет распределение Стюдента с $(n - 1)$ степенью свободы.

Случай а). Определяются двухсторонние критические значения $t_{\alpha/2}$ и $t_{1-\alpha/2}$ как квантили соответственно порядка $\alpha/2$ и порядка $1 - \alpha/2$ распределения Стюдента с $(n - 1)$ степенью свободы. Гипотеза H_0 принимается, если выполняется неравенство $t_{\alpha/2} < T < t_{1-\alpha/2}$, иначе гипотеза H_0 отклоняется.

Случай б). Определяется правостороннее критическое значение t_{α} как квантиль порядка $1 - \alpha$ распределения Стюдента с $(n - 1)$ степенью свободы. Гипотеза H_0 принимается, если $T < t_{\alpha}$.

Случай в). Определяется левостороннее критическое значение t_n как квантиль порядка α распределения Стьюдента с $(n - 1)$ степенью свободы. Гипотеза H_0 принимается, если $t_n \leq T$.

Этот критерий устойчив при умеренных отклонениях распределения выборки от нормального. При проверке равенства в силу симметрии распределения Стьюдента достаточно сравнить $|T|$ с квантилем t_n порядка $1 - \alpha/2$.

Критерий проверки значения дисперсии нормальной совокупности

Статистическая модель. Выборка x_1, x_2, \dots, x_n получена из генеральной совокупности с нормальным законом распределения и с неизвестным математическим ожиданием μ и неизвестной дисперсией σ^2 .

Гипотезы

а) Равенство

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_1: \sigma^2 \neq \sigma_0^2$$

б) Неравенство

$$H_0: \sigma^2 \leq \sigma_0^2$$

$$H_1: \sigma^2 > \sigma_0^2$$

в) Неравенство

$$H_0: \sigma^2 \geq \sigma_0^2$$

$$H_1: \sigma^2 < \sigma_0^2$$

Здесь σ_0^2 — заданное число. Задан уровень значимости α .

Критериальная статистика вычисляется по формуле $T = \frac{(n-1)S_n^2}{\sigma_0^2}$, где

$S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$. При условии истинности гипотезы H_0 статистика T имеет распределение χ^2 с $(n - 1)$ степенью свободы.

Случай а). Определяются двухсторонние критические значения t_n и t_n как квантили соответственно порядка $\alpha/2$ и порядка $1 - \alpha/2$ распределения χ^2 с $(n - 1)$ степенью свободы. Гипотеза H_0 принимается, если выполняется неравенство $t_n \leq T \leq t_n$, иначе гипотеза H_0 отклоняется.

Случай б). Определяется правостороннее критическое значение t_n как квантиль порядка $1 - \alpha$ распределения χ^2 с $(n - 1)$ степенью свободы. Гипотеза H_0 принимается, если $T \leq t_n$.

Случай в). Определяется левостороннее критическое значение t_n как квантиль порядка α распределения χ^2 с $(n - 1)$ степенью свободы. Гипотеза H_0 принимается, если $t_n \leq T$.

Критерий не устойчив, если не выполняется условие нормальности распределения генеральной совокупности.

Случай в). Определяется левостороннее критическое значение t_n как квантиль порядка α распределения Стьюдента с $(n - 1)$ степенью свободы. Гипотеза H_0 принимается, если $t_n < T$.

Этот критерий устойчив при умеренных отклонениях распределения выборки от нормального. При проверке равенства в силу симметрии распределения Стьюдента достаточно сравнить $|T|$ с квантилем t_n порядка $1 - \alpha/2$.

Критерий проверки значения дисперсии нормальной совокупности

Статистическая модель. Выборка x_1, x_2, \dots, x_n получена из генеральной совокупности с нормальным законом распределения и с неизвестным математическим ожиданием μ и неизвестной дисперсией σ^2 .

Гипотезы

а) Равенство

$$H_0: \sigma^2 =$$

$$\sigma_0^2$$

$$H_1: \sigma^2 \neq$$

$$\sigma_0^2$$

б) Неравенство

$$H_0: \sigma^2 <$$

$$\sigma_0^2$$

$$H_1: \sigma^2 >$$

$$\sigma_0^2$$

в) Неравенство

$$H_0: \sigma^2 > \sigma_0^2$$

$$H_1: \sigma^2 < \sigma_0^2$$

(при условии истинности нулевой гипотезы), порядок которых определяется заданным уровнем значимости критерия. На такой основе построены приведенные ниже несколько критериев проверки гипотез о различии между математическими ожиданиями двух нормальных распределений (приведены такие критерии, для которых в Excel предусмотрены специальные средства, описанные в главе 5).

При втором подходе из имеющихся нескольких независимых выборок образуется *единая* общая выборка и критериальная статистика строится на основе общей выборки. Часто такой подход используется при построении непараметрических критериев. Ниже для иллюстрации этого подхода приведен непараметрический критерий Уилкоксона-Манна-Уитни для сравнения распределений двух независимых выборок. Другие критерии этого типа описаны в части III.

Критерий проверки гипотезы о равенстве математических ожиданий для нормальных совокупностей (случай известных дисперсий)

Статистическая модель. Выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_m объемом соответственно n и m извлечены из совокупностей, имеющих нормальные распределения с известными дисперсиями σ_1^2 и σ_2^2 и математическими ожиданиями μ_1 и μ_2 соответственно.

Гипотезы

а) Равенство

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

б) Неравенство

$$H_0: \mu_1 \leq \mu_2$$

$$H_1: \mu_1 > \mu_2$$

Задан уровень значимости α .

По каждой выборке вычисляются выборочные средние $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $\bar{y} = \frac{1}{m} \sum_{i=1}^m y_i$

и затем критериальная статистика $z = \frac{(\bar{x} - \bar{y})}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}$. При условии истинности

гипотезы H_0 статистика z имеет стандартное нормальное распределение.

Случай а). Определяется критическое значение $z_{кр}$ как квантиль порядка $1 - \alpha/2$ стандартного нормального распределения. Гипотеза H_0 принимается, если выполняется неравенство $|z| \leq z_{кр}$, иначе гипотеза H_0 отклоняется.

Случай б). Определяется критическое значение $z_{кр}$ как квантиль порядка $1 - \alpha$ стандартного нормального распределения. Гипотеза H_0 принимается, если $T \leq t$.

Критерий устойчив при умеренных отклонениях распределения выборки от нормального.

В Excel этот критерий реализует средство Двухвыборочный z-тест для средних из пакета анализа (см. раздел 5.6).

Критерий Стьюдента проверки гипотезы о равенстве математических ожиданий для нормальных совокупностей (случай равных дисперсий)

Статистическая модель. Выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_m объемом соответственно n и m извлечены из совокупностей, имеющих нормальные распределения с неизвестными, но равными дисперсиями $\sigma_1^2 = \sigma_2^2 = \sigma^2$ и математическими ожиданиями соответственно μ_1 и μ_2 .

Гипотезы

а) Равенство

б) Неравенство

$$H_0: \mu_1 = \mu_2$$

$$H_0: \mu_1 \leq \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

$$H_1: \mu_1 > \mu_2$$

Задан уровень значимости α .

По каждой выборке вычисляются выборочные средние и выборочные дисперсии: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$, $\bar{y} = \frac{1}{m} \sum_{i=1}^m y_i$, $S_y^2 = \frac{1}{m} \sum_{i=1}^m (y_i - \bar{y})^2$. Поскольку при условии равенства дисперсий имеются две оценки одной и той же величины σ^2 , эти оценки объединяют в одну оценку $S^2 = \frac{(n-1)S_x^2 + (m-1)S_y^2}{(n-1) + (m-1)}$. При условии истинно-

сти гипотезы H_0 величина $T = \frac{\bar{x} - \bar{y}}{S \sqrt{\frac{1}{n} + \frac{1}{m}}}$ имеет распределение Стьюдента с $(n + m - 2)$ степенью свободы. Эта величина T принимается в качестве критериальной статисти-

стики; обычно ее вычисляют по формуле $T = \frac{\sqrt{n+m-2}(\bar{x} - \bar{y})}{\sqrt{\frac{n+m}{nm} [(n-1)S_x^2 + (m-1)S_y^2]}}$, объе-

диняющей две вышеприведенные формулы вычисления S^2 и T .

Случай а). Определяется критическое значение $t_{кр}$ как квантиль порядка $1 - \alpha/2$ распределения Стьюдента с $(n + m - 2)$ степенью свободы. Гипотеза H_0 принимается, если выполняется неравенство $|T| \leq t_{кр}$, иначе гипотеза H_0 отклоняется.

Случай б). Определяется критическое значение $t_{кр}$ как квантиль порядка $1 - \alpha$ распределения Стьюдента с $(n + m - 2)$ степенью свободы. Гипотеза H_0 принимается, если $T \leq t_{кр}$.

Критерий устойчив при умеренных отклонениях распределения выборки от нормального. Критерий также устойчив, если дисперсии генеральных совокупностей незначительно отличаются, а значения n и m приблизительно равны.

В Excel этот критерий реализует средство Двухвыборочный t-тест с одинаковыми дисперсиями из пакета анализа (см. раздел 5.7).

Критерий Стьюдента проверки гипотезы о равенстве математических ожиданий для нормальных совокупностей (случай неравных дисперсий)

Статистическая модель. Выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_m объемом соответственно n и m извлечены из совокупностей, имеющих нормальные распределения с неизвестными дисперсиями σ_1^2 и σ_2^2 и математическими ожиданиями μ_1 и μ_2 соответственно.

Гипотезы

а) Равенство

б) Неравенство

$$H_0: \mu_1 = \mu_2$$

$$H_0: \mu_1 \leq \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

$$H_1: \mu_1 > \mu_2$$

Задан уровень значимости α .

По каждой выборке вычисляются выборочные средние и выборочные дисперсии: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$, $\bar{y} = \frac{1}{m} \sum_{i=1}^m y_i$, $S_y^2 = \frac{1}{m} \sum_{i=1}^m (y_i - \bar{y})^2$. В качестве критериальной статистики берется аналог z-статистики из критерия проверки гипотезы о равенстве средних для нормальных совокупностей при известных дисперсиях: $T = \frac{\bar{x} - \bar{y}}{\sqrt{S_x^2/n + S_y^2/m}}$. Точное распределение этой статистики достаточ-

но сложно, но доказано, что его можно аппроксимировать распределением Стьюдента, если взять число степеней свободы равным

$$k = \frac{(S_x^2/n + S_y^2/m)^2}{\frac{(S_x^2/n)^2}{n-1} + \frac{(S_y^2/m)^2}{m-1}}.$$

Случай а). Определяется критическое значение $t_{кр}$ как квантиль порядка $1 - \alpha/2$ распределения Стьюдента с k степенями свободы. Гипотеза H_0 принимается, если выполняется неравенство $|T| \leq t_{кр}$, иначе гипотеза H_0 отклоняется.

Случай б). Определяется критическое значение $t_{кр}$ как квантиль порядка $1 - \alpha$ распределения Стьюдента с k степенями свободы. Гипотеза H_0 принимается, если $T \leq t_{кр}$.

Этот критерий, если степень свободы распределения Стьюдента вычисляется по приведенной выше формуле, часто называют *критерием Беренса-Фишера*.

Критерий является приближенным. Если нет оснований предполагать, что дисперсии не равны (критерий проверки равенства дисперсий описан ниже), следует применить точный критерий проверки средних при равных дисперсиях. Если сумма объемов выборок больше 30, вместо распределения Стьюдента можно использовать нормальное распределение.

В Excel этот критерий реализует средство Двухвыборочный t-тест с различными дисперсиями из пакета анализа (см. раздел 5.8).

Критерий Стьюдента проверки гипотезы о равенстве математических ожиданий для зависимых нормальных совокупностей

Статистическая модель. Двумерная выборка $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ объемом n извлечена из двумерной нормальной совокупности с неизвестными математическими ожиданиями соответственно μ_1 и μ_2 компонентов выборки.

Гипотезы

а) Равенство

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

б) Неравенство

$$H_0: \mu_1 \leq \mu_2$$

$$H_1: \mu_1 > \mu_2$$

Задан уровень значимости α .

Вычисляются n разностей $d_1 = x_1 - y_1, d_2 = x_2 - y_2, \dots, d_n = x_n - y_n$, и по ним определяются среднее $\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$ и выборочная дисперсия разностей

$$S_d^2 = \frac{1}{n} \sum_{i=1}^n (d_i - \bar{d})^2.$$

Критериальная статистика $T = \frac{\bar{d}}{S_d / \sqrt{n}}$ при условии истинности

нулевой гипотезы имеет распределение Стьюдента с $(n - 1)$ степенью свободы.

Случай а). Определяется критическое значение $t_{кр}$ как квантиль порядка $1 - \alpha/2$ распределения Стьюдента с $(n - 1)$ степенью свободы. Гипотеза H_0 принимается, если выполняется неравенство $|T| \leq t_{кр}$, иначе гипотеза H_0 отклоняется.

Случай б). Определяется критическое значение $t_{кр}$ как квантиль порядка $1 - \alpha$ распределения Стьюдента с $(n - 1)$ степенью свободы. Гипотеза H_0 принимается, если $T \leq t_{кр}$.

В Excel этот критерий реализует средство Парный двухвыборочный t-тест для средних из пакета анализа (см. раздел 5.9).

Непараметрический критерий Уилкоксона–Манна–Уитни для двух независимых выборок

Этот критерий применяется тогда, когда нельзя сделать обоснованных предположений о типе распределений выборок, поскольку он менее мощный, чем аналогичные критерии, основанные на предположениях о конкретных типах распределений генеральных совокупностей.

Критерий Уилкоксона–Манна–Уитни можно применить для проверки гипотезы о равенстве математических ожиданий. Однако заметим, что рассматриваемая нулевая гипотеза, проверяемая с помощью данного критерия, состоит в том, что генеральные совокупности одинаково распределены. Если критерий отклоняет нулевую гипотезу, то это еще не позволяет заключить, что математические ожидания обеих выборок не равны. Для такого вывода необходимо предположить, что рассматриваемые распределения идентичны во всех остальных аспектах, например, что их дисперсии равны. На практике допустимы умеренные различия в значениях дисперсий, так как критерий незначительно чувствителен к ним.

Статистическая модель. Имеются независимые выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_m объемов соответственно n и m . Предполагается, что $n, m \geq 30$.

Гипотезы

H_0 : генеральные совокупности обеих выборок одинаково распределены;

H_1 : нулевая гипотеза неверна.

Задается уровень значимости α .

Для реализации этого критерия выполняются следующие вычисления. Выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_m объединяются в одну общую выборку z_1, z_2, \dots, z_N , $N = n + m$. Значения z_1, z_2, \dots, z_N расставляются в порядке возрастания. Получаем вариационный ряд $z_{(1)} \leq z_{(2)} \leq \dots \leq z_{(N)}$. Номер i местоположения $z_{(i)}$ в этом ряду является рангом данного значения. Ранги принимают значения от 1 до N . Суммируются ранги тех значений, которые принадлежат первой выборке, и получается число R_1 . Аналогично определяется R_2 — сумма рангов второй выборки. Если два (или более) выборочных значения имеют одинаковые значения, то каждому из них приписывается значение ранга, равное среднему из рангов, которые были бы им приписаны при отсутствии совпадений.

Вычисляются величины $U_1 = nm + \frac{n(n+1)}{2} - R_1$ и $U_2 = nm + \frac{m(m+1)}{2} - R_2$, из которых выбирается наибольшая, т.е. $U = \max(U_1, U_2)$. Вычисляется критериаль-

ная статистика $T = \frac{U - \frac{1}{2}nm}{\sqrt{\frac{nm(N+1)}{12}}}$. При условии истинности гипотезы H_0 статистика

T имеет асимптотически стандартное нормальное распределение.

Вычисляется критическое значение $t_{кр}$ как квантиль порядка $1 - \alpha/2$ стандартного нормального распределения. Гипотеза H_0 принимается, если выполняется неравенство $|T| \leq t_{кр}$, иначе гипотеза H_0 отклоняется.

Приведенная критериальная статистика T применяется при больших выборках. При малых выборках в качестве критериальной статистики используется величина U , а критическая область определяется по специальному распределению Манна-Уитни.

Критерий Фишера проверки равенства дисперсий двух независимых выборок из нормально распределенных генеральных совокупностей

Статистическая модель. Выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_m объемом соответственно n и m извлечены из совокупностей, имеющих нормальные распределения с неизвестными дисперсиями σ_1^2 и σ_2^2 и математическими ожиданиями μ_1 и μ_2 соответственно.

Гипотезы

а) Равенство

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

б) Неравенство

$$H_0: \sigma_1^2 \leq \sigma_2^2$$

$$H_1: \sigma_1^2 > \sigma_2^2$$

Задан уровень значимости α .

Для каждой выборки вычисляются выборочные дисперсии $S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$,

$$S_y^2 = \frac{1}{m} \sum_{i=1}^m (y_i - \bar{y})^2 \text{ и их отношение } F = \frac{S_x^2}{S_y^2}. \text{ Это отношение, называемое дисперсионным отношением Фишера, выбирается в качестве критериальной статистики}$$

и в случае истинности нулевой гипотезы имеет F -распределение со степенями свободы $k_1 = n - 1$ и $k_2 = m - 1$ (о F -распределении речь идет в разделе 1.5.7).

Случай а). Определяется критическое значение $t_{кр}$ как квантиль порядка $1 - \alpha/2$ F -распределения со степенями свободы $k_1 = n - 1$ и $k_2 = m - 1$. Гипотеза H_0 принимается, если выполняется неравенство $F \leq t_{кр}$, иначе гипотеза H_0 отклоняется.

Случай б). Определяется критическое значение $t_{кр}$ как квантиль порядка $1 - \alpha$ F -распределения со степенями свободы $k_1 = n - 1$ и $k_2 = m - 1$. Гипотеза H_0 принимается, если $F \leq t_{кр}$.

В Excel этот критерий реализует средство Двухвыборочный F -тест для дисперсий из пакета анализа (см. раздел 5.10).

нал статистика $G = \frac{\frac{1}{jnm} \sum_{i=1}^j \sum_{k=1}^m (x_{ik} - \bar{x}_{i.})^2}{\frac{1}{nm} \sum_{i=1}^j \sum_{k=1}^m (x_{ik} - \bar{x}_{.k})^2}$. При условии истинности гипотезы H_0 статистика

T имеет асимптотически стандартное нормальное распределение.

Вычисляется критическое значение $t_{кр}$ как квантиль порядка $1 - \alpha/2$ стандартного нормального распределения. Гипотеза H_0 принимается, если выполняется неравенство $|T| \leq t_{кр}$, иначе гипотеза H_0 отклоняется.

Приведенная критериальная статистика T применяется при больших выборках. При малых выборках в качестве критериальной статистики используется

конкретному классу распределении сложно построить достаточно точные доверительные интервалы для оцениваемых параметров распределения или подобрать наиболее мощный критерий проверки сформулированных гипотез.

Рассмотрим два критерия данного типа: критерий χ^2 (также называемый критерием согласия, критерием согласия χ^2 или критерием Пирсона) и критерий Колмогорова. Первый из них является более общим, но, как правило, менее точным, чем второй. С другой стороны, критерий Колмогорова применяется только для непрерывных распределений.

Критерий χ^2

Рассмотрим последовательность независимых испытаний, в каждом из которых может осуществиться один из исходов A_1, A_2, \dots, A_m с вероятностями p_1, p_2, \dots, p_m соответственно ($\sum_{i=1}^m p_i = 1$). Пусть проведено n испытаний, при этом событие A_1 наблюдалось v_1 раз, событие A_2 наблюдалось v_2 раз и т.д., событие A_m наблюдалось v_m раз ($\sum_{i=1}^m v_i = n$). Распределение случайной величины

$$\eta = \sum_{i=1}^m \frac{(v_i - np_i)^2}{np_i} = \sum_{i=1}^m \frac{v_i^2}{np_i} - n$$

при $n \rightarrow \infty$ стремится к распределению χ^2 с $(m - 1)$ степенью свободы (теорема К. Пирсона).

Это свойство случайной величины η позволяет взять ее в качестве критериальной статистики для критерия проверки гипотез о принадлежности распределения выборки классу распределений. Рассмотрим этот критерий для непрерывных распределений (все варианты критерия и их практическая реализация приведены в главе 9, раздел 9.2).

Статистическая модель. Выборка, состоящая из независимых выборочных значений x_1, x_2, \dots, x_n , получена из генеральной совокупности, имеющей функцию распределения $F(u)$, зависящей от k параметров, из которых k_1 параметров неизвестно.

Гипотезы

H_0 : выборочные значения получены из генеральной совокупности с функцией распределения $F(u)$, зависящей от k параметров, из которых k_1 параметров определяются по выборочным значениям;

H_1 : нулевая гипотеза неверна.

Задается уровень значимости α .

Чтобы построить критериальную статистику, область возможных выборочных значений разбивается на m непересекающихся интервалов $\Delta_1 = (x^{(1)}, x^{(2)})$, $\Delta_2 = (x^{(2)}, x^{(3)})$, ..., $\Delta_m = (x^{(m)}, x^{(m+1)})$. Подсчитывается, сколько выборочных значений попало в каждый интервал Δ_i . Получается ряд частот v_1, v_2, \dots, v_m (при этом, конечно, должно выполняться равенство $v_1 + v_2 + \dots + v_m = n$, где n — объем выборки). В предположении, что справедлива гипотеза H_0 , по формуле $p_i = np_i = n[F(x^{(i+1)}) - F(x^{(i)})]$ вычисляются ожидаемые значения частот, т.е. количество попаданий выборочных значений в каждый из интервалов Δ_i , где $x^{(i)}$ и $x^{(i+1)}$ — границы интервала Δ_i . Теперь можно вычислить критериальную статистику

$T = \sum_{i=1}^n \frac{(v_i - np_i)^2}{np_i}$. Отметим, что, поскольку k_1 параметров распределения определяется на основе выборочных значений, распределение χ^2 , которое асимптотически имеет статистика T , имеет $(m - k_1 - 1)$ степеней свободы.

Критическое значение критерия $t_{кр}$ определяется как квантиль порядка $1 - \alpha$ распределения χ^2 с $(m - k_1 - 1)$ степенью свободы. Гипотеза H_0 принимается, если $T \leq t_{кр}$. В противном случае гипотеза H_0 отклоняется.

Проблема выбора количества и построение интервалов $\Delta_i = (x^{(i)}, x^{(i+1)})$ рассмотрены в разделе 9.2.

Критерий Колмогорова

Статистическая модель. Выборка, состоящая из независимых выборочных значений x_1, x_2, \dots, x_n , получена из генеральной совокупности, распределение которой предполагается непрерывным.

Гипотезы

H_0 : выборочные значения получены из генеральной совокупности с функцией распределения $F(u)$;

H_1 : нулевая гипотеза неверна.

Задается уровень значимости α .

Критериальная статистика здесь определяется как максимум отклонения выборочного распределения $F_n(u)$ (строится по выборке) от гипотетической функции распределения $F(u)$. Для вычисления такой статистики выполняются следующие действия.

По выборке x_1, x_2, \dots, x_n строится вариационный ряд $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. Затем вычисляются так называемые кумулятивные разности: $D_m^+ = \frac{m}{n} - F(x_{(m)})$

и $D_m^- = F(x_{(m)}) - \frac{m-1}{n}$, $m = 1, 2, \dots, n$. После вычисляется критериальная статистика $D_n = \max_{1 \leq m \leq n} (D_m^+, D_m^-)$. При условии истинности гипотезы H_0 статистика D_n имеет так называемое распределение Колмогорова–Смирнова.

Критическое значение $t_{кр}$ определяется как квантиль порядка $1 - \alpha$ распределения Колмогорова–Смирнова. Гипотеза H_0 принимается, если $D_n \leq t_{кр}$. В противном случае гипотеза H_0 отклоняется.

Практическая реализация этого критерия показана в главе 9 (раздел 9.3).

Анализ статистических зависимостей

В этой главе рассмотрены задачи и методы анализа статистических зависимостей, которые включают в себя широкий спектр статистических алгоритмов. Но прежде чем перейти к формулировке общей и частных задач статистического анализа зависимостей, представим весьма общую модель, в рамках которой легче понять и сформулировать эти задачи.

3.1. Общая модель статистических зависимостей

Большое количество природных явлений, явлений общественной жизни, моделирование технических устройств, технологических процессов и т.п. можно представить в виде следующей математической модели, которая будет описывать все эти разнородные явления и процессы. Есть некоторый объект (система, процесс, явление и т.д.), на входе которого наблюдается *“входное” воздействие* X , а на выходе — *результатирующая переменная* Y . Существует также *случайное воздействие* ϵ на объект, не поддающееся непосредственному измерению и контролю. В общем виде такая схема представлена на рис. 3.1. Переменные X , Y , ϵ в общем случае являются векторными переменными различных размерностей, т.е. $X = (x^{(1)}, x^{(2)}, \dots, x^{(p)})$, $Y = (y^{(1)}, y^{(2)}, \dots, y^{(m)})$, $\epsilon = (\epsilon^{(1)}, \epsilon^{(2)}, \dots, \epsilon^{(k)})$, при этом все или некоторые компоненты векторов X , Y и ϵ могут быть функциями от времени (временными процессами).



Рис. 3.1. Общее представление объекта со случайным воздействием

Входная переменная X описывает условия функционирования объекта (часть компонентов вектора X , как правило, поддается регулированию или частичному управлению); в различных статистических моделях компоненты вектора X называют

независимыми переменными, фактор-аргументами, экзогенными, предикторными (или просто предикторами, т.е. предсказывателями), объясняющими и т.д.

Компоненты $y^{(1)}, y^{(2)}, \dots, y^{(m)}$ вектора Y — это *выходные* переменные, характеризующие *поведение* или *результат* функционирования объекта; в статистических моделях их называют зависимыми, откликами, эндогенными, результирующими или объясняемыми переменными.

Компоненты $e^{(1)}, e^{(2)}, \dots, e^{(m)}$ случайного вектора e — это латентные (т.е. скрытые) случайные "остаточные" компоненты, отражающие влияние на Y неучтенных "на входе" факторов, а также случайные ошибки в измерении анализируемых показателей.

Среди компонентов векторов X и Y могут быть переменные следующих типов.

- *Количественные*, т.е. принимающие числовые значения, измеренные в определенной шкале (например, денежный доход и сбережения семьи в социологии, численность популяции и линейные размеры особи в биологии, потребляемая энергия и выходная мощность в технике и т.п.).
- *Порядковые* (или *ординальные*), т.е. позволяющие упорядочивать анализируемые объекты по степени проявления в них изучаемого свойства (уровень образования работников или уровень жилищных условий в социологии, степень какого-либо заболевания в медицине и т.п.).
- *Классификационные* (или *номинальные*), позволяющие разбивать совокупность объектов на не поддающиеся упорядочению однородные по анализируемому свойству классы (профессия работника, мотивы миграции в социологии, пол особи, вид и род в биологии и т.д.).

Отметим, что тип переменных существенно влияет на выбор применяемых статистических методов.

3.2. Задачи статистического анализа зависимостей

Общая задача статистического анализа зависимостей может быть сформулирована следующим образом:

по результатам n измерений $(X_1, Y_1), (X_2, Y_2), \dots, (X_n, Y_n)$ исследуемых переменных X и Y построить такую функцию $f(X)$ (в общем случае X и Y являются векторами, функция $f(X)$ — векторозначная), которая позволила бы наилучшим образом, в определенном смысле, восстанавливать значения результирующих переменных $Y = (y^{(1)}, y^{(2)}, \dots, y^{(m)})$ по заданным значениям входных переменных $X = (x^{(1)}, x^{(2)}, \dots, x^{(n)})$.

Данная формулировка задачи нуждается в уточнениях. В частности, прежде всего необходимо ответить на следующие вопросы.

- Каково математическое выражение искомой зависимости между X и Y , записанной в терминах $Y, X, f(X)$ и e ?
- В соответствии с каким *критерием качества* аппроксимации будет определяться наилучший способ восстановления значений Y ?

- С какой *прикладной целью* проводится статистический анализ, т.е. для решения каких *конкретных задач* будет использована построенная в результате анализа функция $f(X)$?

С последнего вопроса должен начинаться любой статистический анализ зависимостей — от ответа на этот вопрос существенно зависят последовательность выполнения различных этапов анализа, выбор общей структуры функции f , интерпретация полученных статистических результатов и т.д.

Выделим три основных типа конечных прикладных целей (задач) анализа зависимостей, расположив их как бы по нарастанию степени проникновения в содержательную сущность анализируемой конкретной задачи.

Тип 1. Установление самого факта наличия (или отсутствия) статистически значимой связи между Y и X . Выбор вида функции f играет подчиненную роль, и часто даже не стоит вопрос о построении функции f . Задачи этого типа решаются методами корреляционного анализа, ранговых корреляций и с помощью анализа таблиц сопряжения.

Тип 2. Прогноз (восстановление) значений результирующих переменных Y по заданным значениям выходных переменных X . Здесь также выбор функции f играет подчиненную роль, поскольку в данном случае интересуются лишь значениями функции $f(X)$, но не ее структурой, т.е. функция f должна хорошо аппроксимировать “числовую” зависимость между Y и X , но совсем не обязана отражать “физическую” связь между X и Y .

Тип 3. Выявление причинных связей между входными переменными X и результирующими переменными Y . Такая постановка задачи претендует на проникновение в “физический механизм” изучаемых статистических связей, т.е. в тот самый механизм преобразования входных переменных X и ϵ в результирующие показатели Y . Здесь на первый план выходит задача правильного определения структуры функции $f(X)$, при этом часто параметры, от которых может зависеть функция f , имеют определенную “физическую” интерпретацию.

Задачи типов 2 и 3 решаются методами регрессионного и дисперсионного анализа, дискриминантного анализа и др.

Приведем таблицу статистических методов, которые “обслуживают” тот или иной тип задач в зависимости от природы изучаемых переменных.

<i>Вид результирующих переменных Y</i>	<i>Вид входных переменных X</i>	<i>Разделы статистического анализа</i>
Количественные	Количественные	Регрессионный и корреляционный анализ
Количественные	Единственная количественная переменная, интерпретируемая как “время”	Анализ временных рядов
Количественные	Неколичественные (порядковые или классификационные)	Дисперсионный анализ
Количественные	Смешанные (количественные и неколичественные)	Ковариационный анализ, модели типологической регрессии

<i>Вид результирующих переменных Y</i>	<i>Вид входных переменных X</i>	<i>Разделы статистического анализа</i>
Неколичественные (порядковые или классификационные)	Неколичественные (порядковые или классификационные)	Анализ ранговых корреляций и таблиц сопряжения
Неколичественные (порядковые или классификационные)	Количественные	Дискриминантный анализ, кластер-анализ, таксономия, расщепление смесей распределений
Смешанные (количественные и неколичественные)	Смешанные (количественные и неколичественные)	Аппарат логических решающих функций

Опишем более подробно основные задачи анализа статистических зависимостей и методы их решения.

3.3. Корреляционный анализ

В этом разделе рассмотрим задачу установления самого факта наличия статистически значимой связи между переменными. Это задача типа I из предыдущего раздела. Методы, применяемые для ее решения, зависят от природы исследуемых случайных переменных (количественные, порядковые или классификационные), от выбранного показателя статистической зависимости (индекс или коэффициент корреляции, ранговый коэффициент корреляции и т.п.) и от конкретной решаемой задачи: точечное и/или интервальное оценивание показателя статистической зависимости, проверка гипотезы о значении показателя статистической зависимости (как правило, проверяется гипотеза о статистически значимом отличии этого показателя от нуля). Конечно, могут быть поставлены и другие конкретные задачи, например установление структуры связей между компонентами входной переменной X и выходной переменной Y.

Опишем возможные методы решения перечисленных задач в зависимости от вида исследуемых случайных переменных.

3.3.1. Анализ зависимостей между количественными переменными

Представим два показателя статистической зависимости (коэффициент корреляции и индекс корреляции), используемых для анализа статистической зависимости между количественными переменными.

Коэффициент корреляции

Пусть анализируется парная зависимость между случайными переменными X и Y. Напомним (см. раздел 1.2.5), что коэффициент корреляции между случайными величинами X и Y определяется как $\rho = \frac{\text{cov}(X, Y)}{\sqrt{DX \cdot DY}}$, где ковариация $\text{cov}(X, Y)$ вычисляется по формуле $\text{cov}(X, Y) = M[(X - MX)(Y - MY)]$. Значение

коэффициента корреляции лежит между -1 и 1 . Он характеризует степень *линейной* зависимости между величинами X и Y .

Приведем точечные и интервальные оценки выборочного коэффициента корреляции.

Точечные оценки

Статистическая модель. Выборочные значения $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ являются реализацией случайной величины $Z = (X, Y)$, имеющей произвольное двумерное распределение с конечными моментами второго порядка.

Ниже приведена статистика для оценки коэффициента корреляции:

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}},$$

$$\text{где } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

Распределение этой статистики в общем случае достаточно сложно и зависит от неизвестного значения коэффициента корреляции ρ . Для выборочного коэффициента корреляции известно нормализующее *z-преобразование Фишера*

$z = \frac{1}{2} \ln \frac{1+r}{1-r}$, замечательное тем, что распределение случайной величины z не за-

висит от неизвестного коэффициента корреляции. Кроме того, уже при $n \geq 20$ это распределение близко к нормальному, причем

$$Mz = \frac{1}{2} \ln \frac{1+r}{1-r} + \frac{r}{2(n-3)} \left[1 - \frac{3-r^2}{4(n-3)} + \dots \right],$$

$$Dz = \frac{1}{n-3} \left[1 - \frac{r^2}{2(n-3)} - \frac{2-6r^2+3r^4}{6(n-3)^2} + \dots \right],$$

$$\beta_1(z) = \frac{[M(z - Mz)^3]^2}{(Dz)^3} = \frac{r^6}{(n-3)^2} + \dots,$$

$$\beta_2(z) = \frac{M(z - Mz)^4}{(Dz)^2} = 3 + \frac{2}{n-3} + \frac{2r^2 - 3r^4}{(n-3)^2} + \dots$$

Здесь $\beta_1(z)$ — коэффициент асимметрии, $\beta_2(z)$ — коэффициент эксцесса случайной величины z . Отметим, что при вычислении математического ожидания и дисперсии случайной величины z в приведенных выше формулах обычно ограничиваются

лишь первыми слагаемыми, т.е. полагают, что $Mz = \frac{1}{2} \ln \frac{1+r}{1-r}$ и $Dz = \frac{1}{n-3}$.

Интервальные оценки для коэффициента корреляции

Поскольку случайная величина $(z - Mz)/\sqrt{Dz}$ распределена приблизительно по стандартному нормальному закону, это свойство используется при построении доверительного интервала для коэффициента корреляции. Если

задан доверительный уровень α , из уравнения $\alpha = 2\Phi(k) - 1$, где Φ — функция распределения стандартного нормального закона, определяется коэффициент k . Затем вычисляются границы z_1 и z_2 доверительного интервала для z :

$$z_1 = \frac{1}{2} \ln \frac{1+r}{1-r} - \frac{k}{\sqrt{n-3}} \quad \text{и} \quad z_2 = \frac{1}{2} \ln \frac{1+r}{1-r} + \frac{k}{\sqrt{n-3}}.$$

Отсюда в результате обратного преобразования Фишера вычисляются границы r_1 и r_2 доверительного интервала (r_1, r_2) для коэффициента корреляции:

$$r_1 = \frac{e^{2z_1} - 1}{e^{2z_1} + 1} \quad \text{и} \quad r_2 = \frac{e^{2z_2} - 1}{e^{2z_2} + 1}.$$

Практическая реализация этого метода построения доверительного интервала показана в главе 13. В этой же главе приведены критерии проверки гипотез о значении коэффициента корреляции.

Индекс корреляции и коэффициент детерминации

Индекс корреляции применяется в модели $Y(X) = f(X) + \epsilon$, где ϵ — случайная переменная, а переменная X может быть вектором. Таким образом, индекс корреляции можно применять там, где не применим “стандартный” коэффициент корреляции, используемый для анализа парных наблюдений.

Обозначим через σ_Y^2 общую дисперсию случайной величины Y , через σ_f^2 — дисперсию функции $f(X)$, а через σ_ϵ^2 — остаточную дисперсию, определяемую случайной величиной ϵ (формулы для вычисления этих дисперсий приведены в разделе 3.4.3). Эти три дисперсии связаны равенством $\sigma_Y^2 = \sigma_f^2 + \sigma_\epsilon^2$.

Индексом корреляции I_{YX} называется величина, определяемая соотношением $I_{YX}^2 = \frac{\sigma_f^2}{\sigma_Y^2} = 1 - \frac{\sigma_\epsilon^2}{\sigma_Y^2}$. Очевидно, что $0 \leq I_{YX} \leq 1$. Если $I_{YX} = 0$, тогда $\sigma_f^2 = 0$ или, что то же самое, $\sigma_Y^2 = \sigma_\epsilon^2$. Это означает полное отсутствие какого-либо влияния переменной X на переменную Y , т.е. отсутствие корреляционной связи между X и Y . Если же $I_{YX} = 1$, то $\sigma_\epsilon^2 = 0$. Это означает наличие чисто функциональной зависимости между переменными X и Y .

Квадрат индекса корреляции показывает, какая доля дисперсии результирующей величины Y определяется (детерминируется) вариацией (дисперсией) функции $f(X)$, зависящей от влияющей переменной X . Поэтому квадрат индекса корреляции часто называют коэффициентом детерминации и обозначают как R^2 . Этот коэффициент используется как мера адекватности подбора функции регрессии для аппроксимации исходных данных (см. раздел 3.4.3).

3.3.2. Анализ зависимостей между порядковыми переменными

Напомним, что порядковыми (ординальными) называют величины, значения которых можно ранжировать в соответствии с некоторой заданной шкалой. Таким образом, значениями подобных величин считаются ранги, присвоенные им в соответствии с этой шкалой. Классическими примерами таких величин являются уровень образования работников в социологии, уровень использования высоких технологий в промышленности какого-либо региона или страны в целом,

степень эффективности некоего медицинского препарата для лечения ряда заболеваний и т.п. Количественные величины являются частным случаем порядковых. Сгруппированные количественные величины также можно рассматривать как порядковые величины. (Именно поэтому ранговые коэффициенты корреляции, описанные ниже, часто применяются и для анализа зависимостей между количественными переменными.)

Если для анализа предоставляется выборка, не преобразованная в ранги, то сначала эту выборку необходимо преобразовать следующим образом. Пусть наблюдается двумерная случайная величина $Z = (X, Y)$. В результате имеем выборку объемом n $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Каждому выборочному значению (x_i, y_i) присваиваются ранги (r_i, q_i) . Таким образом, вместо исходной выборки имеем совокупность двумерных значений $(r_1, q_1), (r_2, q_2), \dots, (r_n, q_n)$. Ранги присваиваются значениям x_i и y_i независимо путем построения отдельных вариационных рядов $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ и $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$ (см. раздел 2.3.9). Число i члена вариационного ряда $x_{(i)}$ будет рангом соответствующего выборочного значения. Если есть совпадающие выборочные значения, то им присваиваются одинаковые ранги, равные среднему рангов, которые были бы им присвоены при отсутствии равенства значений. Например, пусть значения $x_{(k)}, x_{(k+1)}$ и $x_{(k+2)}$ равны между собой, тогда они получают один и тот же ранг $(k + k + 1 + k + 2)/3 = k + 1$. Поэтому некоторые ранги могут быть дробными. Далее будем предполагать, что выборочные значения $(r_1, q_1), (r_2, q_2), \dots, (r_n, q_n)$ являются рангами.

Для оценивания степени зависимости между порядковыми случайными величинами разработаны специальные ранговые коэффициенты корреляции. На практике наиболее часто используется ранговый коэффициент корреляции Спирмена.

Ранговый коэффициент корреляции Спирмена

Этот коэффициент корреляции вычисляется по формуле

$$r_s = 1 - \frac{6}{n^3 - n} \sum_{i=1}^n (r_i - q_i)^2.$$

Доказано, что коэффициент корреляции Спирмена по модулю не превосходит 1 (так же, как и обычный коэффициент корреляции). Если все ранги (r_i, q_i) попарно совпадают, то $r_s = 1$. Если же эти ранги *противоположны*, т.е. $q_i = n - r_i + 1$, то $r_s = -1$. Отметим, что, если некоторые ранги совпадают, существует своя значительно более сложная формула вычисления коэффициента корреляции, но на практике и в этом случае используют вышеприведенную формулу.

При условии независимости случайных величин X и Y $M(r_s) = 0$ и $D(r_s) = 1/(n - 1)$. Для количественных случайных величин коэффициент корреляции Спирмена близок к обычному коэффициенту корреляции. Например, в случае двумерной нормально распределенной случайной величины $Z = (X, Y)$, для компонентов которой коэффициент корреляции равен ρ , соотношение между коэффициентом корреляции Спирмена r_s и коэффициентом корреляции ρ имеет

$$\text{вид } r_s = \frac{6}{\pi} \arcsin \frac{\rho}{2} = \frac{3}{\pi} \left(\rho + \frac{\rho^3}{24} + \frac{3\rho^5}{640} + \dots \right).$$

Ранговый коэффициент корреляции Спирмена обычно применяется для проверки гипотезы о зависимости или независимости случайных величин X и Y . Для проверки этой гипотезы по малым выборкам ($n \leq 10$) в качестве

критериальной статистики используется коэффициент r_s , а критические значения при заданном уровне значимости определяются по таблицам распределения этого коэффициента. Для больших выборок ($n > 10$) в качестве критериальной статистики берется величина $t = \frac{r_s \sqrt{n-2}}{\sqrt{1-r_s^2}}$, которая асимптотически имеет распределение Стьюдента с $(n - 2)$ степенью свободы.

Ранговый коэффициент корреляции Кендалла

“Конкурентом” коэффициенту корреляции Спирмена для оценивания степени зависимости между порядковыми случайными величинами может служить ранговый коэффициент корреляции Кендалла.

Пусть для выборочных значений $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ вычислены ранги $(r_1, q_1), (r_2, q_2), \dots, (r_n, q_n)$. Эта последовательность упорядочивается по возрастанию рангов r_i , и получается последовательность $(1, q_{(1)}), (2, q_{(2)}), \dots, (n, q_{(n)})$. Ранговый коэффициент корреляции Кендалла вычисляется по формуле

$$r_K = \frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j=i+1}^n \text{sign}(q_{(i)} - q_{(j)}),$$

где функция $\text{sign}(x)$ принимает значение $+1$, если $x > 0$, и значение -1 , если $x < 0$. Коэффициент корреляции Кендалла по модулю не превосходит 1 и при условии независимости случайных величин X и Y $M(r_K) = 0$ и $D(r_K) = \frac{2(2n+5)}{9n(n-1)}$.

Ранговый коэффициент корреляции Кендалла (так же, как и коэффициент Спирмена) применяется для проверки гипотезы о зависимости или независимости случайных величин X и Y . Для проверки этой гипотезы по малым выборкам ($n \leq 10$) в качестве критериальной статистики используется коэффициент r_K , а критические значения при заданном уровне значимости определяются по таблицам распределения этого коэффициента. Для больших выборок ($n > 10$) в качестве критериальной статистики берется величина $t = r_K \sqrt{\frac{9n(n-1)}{2(2n+5)}}$, которая асимптотически имеет стандартное нормальное распределение.

Если сравнивать применение коэффициентов корреляции Спирмена и Кендалла для проверки гипотезы о зависимости или независимости случайных величин X и Y , то считается, что коэффициент Кендалла дает более точные результаты, особенно для малых выборок. Кроме того, построение доверительных интервалов для неизвестных истинных значений ранговых коэффициентов корреляции возможно только приближенно и только на основе коэффициента Кендалла.

Коэффициент согласованности множественных связей

Ранговые коэффициенты корреляции Спирмена и Кендалла применяются для оценки статистических связей между двумя порядковыми переменными. Иногда возникает необходимость в оценке статистической зависимости между несколькими (больше двух) переменными. Для этих целей используется коэффициент согласованности (также называемый коэффициентом конкордации).

Пусть наблюдается m -мерная случайная величина $Z = (X_1, X_2, \dots, X_m)$. В результате имеем выборку объемом n $(x_{11}, x_{21}, \dots, x_{m1}), (x_{12}, x_{22}, \dots, x_{m2}), \dots$,

$(x_{1n}, x_{2n}, \dots, x_{mn})$. Каждому выборочному значению $(x_{1i}, x_{2i}, \dots, x_{mi})$ присваиваются ранги $(r_{1i}, r_{2i}, \dots, r_{mi})$. Ранги r_{ji} присваиваются значениям x_{ji} независимо путем построения отдельных вариационных рядов для реализации каждого компонента X_j так же, как при вычислении коэффициентов Спирмена и Кендалла. Если есть совпадающие выборочные значения, то им присваиваются одинаковые ранги, равные среднему рангов, которые были бы им присвоены при отсутствии равенства значений.

Коэффициент согласованности вычисляется по формуле

$$W = \frac{12}{m^2(n^3 - n)} \sum_{i=1}^n \left(\sum_{j=1}^m r_{ji} - \frac{m(n+1)}{2} \right)^2.$$

Этот коэффициент принимает значения из интервала $[0, 1]$. Если $W = 0$, то считается, что компоненты X_1, X_2, \dots, X_m независимы. С другой стороны, $W = 1$ тогда и только тогда, когда все ранги r_{ji} , соответствующие выборочному значению $(x_{1i}, x_{2i}, \dots, x_{mi})$, равны и это условие выполняется для всех выборочных значений. При условии независимости случайных величин X_1, X_2, \dots, X_m $M(W) = 1/m$

и $D(W) = \frac{2(n-1)}{m^3(n-1)}$. Отметим, что при $m=2$ $W \approx (1 + r_s)/2$, где r_s — коэффициент корреляции Спирмена.

Для проверки гипотезы о зависимости или независимости случайных величин X_1, X_2, \dots, X_m по малым выборкам в качестве критериальной статистики используется коэффициент W , а критические значения при заданном уровне значимости определяются по таблицам распределения этого коэффициента. Данное распределение удовлетворительно аппроксимируется бета-распределением [4]. Для выборок объемом более 7 в качестве критериальной статистики берется величина $t = m(n-1)W$, которая асимптотически имеет распределение χ^2 с $(n-1)$ степенью

свободы. Иногда используется статистика $T = \frac{1}{2} \ln \left(\frac{(m-1)W}{1-W} \right)$, которая приближенно

имеет F -распределение со степенями свободы $v_1 = n-1-2/m$ и $v_2 = (m-1)v_1$.

Практическая реализация описанных ранговых коэффициентов корреляции показана в главе 13.

3.3.3. Анализ зависимостей между классификационными переменными

Напомним, что классификационные (номинальные) переменные принимают значения, которые можно разбить на непересекающиеся множества, но эти множества трудно или невозможно упорядочить по какому-либо признаку. "Классическими" примерами таких переменных являются профессии работников или мотивы миграции в социологии, пол особи, вид и род в биологии и т.д. Если хотя бы одна из переменных является количественной, такие данные исследуются методами дисперсионного анализа (в этом случае неколичественные переменные можно отождествить с факторами влияния; см. раздел 3.5). В общем случае основным инструментом исследования зависимостей между классификационными переменными являются *таблицы сопряженности*.

Рассмотрим двумерные таблицы сопряженности, которые соответствуют двум классификационным переменным (такие таблицы иногда называют таблицами

сопряженности с двумя входами). Анализ многомерных таблиц сопряженности (таблиц с тремя и более входами) достаточно сложен; методы анализа таких таблиц можно найти в [3].

Пусть имеется двумерная случайная величина $Z = (X, Y)$, где случайная величина X принимает значения (признаки) A_1, A_2, \dots, A_s , а случайная величина Y — значения (признаки) B_1, B_2, \dots, B_r .¹ Выборочные данные представляются в виде следующей таблицы сопряженности. Здесь x_{ij} — количество выборочных значений, имеющих признаки B_i и A_j .

	A_1	A_2	...	A_s	Всего
B_1	x_{11}	x_{12}	...	x_{1s}	$n_{1\cdot} = \sum_{j=1}^s x_{1j}$
B_2	x_{21}	x_{22}	...	x_{2s}	$n_{2\cdot} = \sum_{j=1}^s x_{2j}$
...
B_r	x_{r1}	x_{r2}	...	x_{rs}	$n_{r\cdot} = \sum_{j=1}^s x_{rj}$
Всего	$n_{\cdot 1} = \sum_{i=1}^r x_{i1}$	$n_{\cdot 2} = \sum_{i=1}^r x_{i2}$...	$n_{\cdot s} = \sum_{i=1}^r x_{is}$	$n = \sum_{i=1}^r n_{i\cdot} = \sum_{j=1}^s n_{\cdot j}$

Для проверки гипотезы о независимости случайных величин X и Y вычисляется критерияльная статистика

$$T = n \sum_{i=1}^r \sum_{j=1}^s \frac{(x_{ij} - n_{i\cdot} n_{\cdot j})^2}{n_{i\cdot} n_{\cdot j}} = n \left(\sum_{i=1}^r \sum_{j=1}^s \frac{x_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 1 \right).$$

Эта статистика приближенно имеет распределение χ^2 со степенью свободы, равной $(r-1)(s-1)$. Для случая $r=s=2$ имеется точный критерий Фишера проверки гипотезы о независимости [14, 17].

Если критерий проверки гипотезы о независимости устанавливает, что существует статистически значимая зависимость между переменными X и Y , то полезно иметь какую-то числовую меру этой зависимости (наподобие коэффициента корреляции для количественных переменных). Статистика T в силу ряда причин непосредственно не может выступать в качестве такой меры зависимости, однако на ее основе разработано несколько показателей зависимости классификационных переменных, среди которых выделим следующие:

¹ Поясняющий пример. Пусть необходимо проверить, есть ли зависимость между цветом глаз и цветом волос у людей (пример из [14]). Если случайная величина X — это "цвет глаз", а величина Y — "цвет волос", тогда A_1 = "карий цвет глаз", A_2 = "синий цвет глаз" и т.д., B_1 = "блондин(ка)", B_2 = "брюнет(ка)" и т.д. Каждый индивидум, информация о котором включена в исследуемую выборку, характеризуется двумя признаками: A_i и B_k , где i — номер индивидуума, j — номер цвета глаз, k — номер цвета волос.

- коэффициент сопряженности $C = \sqrt{\frac{T}{T+n}}$;
- мера связи Чупрова $K = \sqrt{\frac{T}{n\sqrt{(r-1)(s-1)}}}$;
- коэффициент $\varphi = \sqrt{\frac{T}{n}}$.

Эти коэффициенты используются в различных ситуациях и каждый из них имеет свои преимущества и недостатки.

В заключение отметим, что для анализа зависимости классификационных переменных разработаны так называемые *информационные показатели зависимости*, использующие понятие энтропии и количества информации, что позволяет определять *направленные меры зависимости* между переменными. Эти весьма интересные показатели зависимости описаны в [1].

3.4. Регрессионный анализ

Рассмотрим более подробно виды зависимостей между количественными переменными X и Y (одна или обе эти переменные могут быть векторными). Здесь возможны следующие случаи.

Регрессионная зависимость случайного результирующего показателя Y от неслучайных входных переменных X . Природа такой связи может носить двойственный характер:

- а) регистрация результирующего показателя Y неизбежно связана с некоторыми случайными ошибками измерения e , в то время как входные переменные X измеряются без ошибок (или величины этих ошибок пренебрежимо малы по сравнению с ошибками измерения Y);
- б) значения переменных Y зависят не только от соответствующих значений X , но и от ряда неконтролируемых факторов, поэтому при каждом фиксированном значении X соответствующие значения результирующего показателя $Y(X)$, измеренные в ряде опытов, неизбежно подвержены некоторому случайному разбросу.

Удобной математической моделью такого рода зависимостей является уравнение вида $Y(X) = f(X) + \epsilon$, где ϵ — случайная переменная. Это уравнение называется *уравнением регрессии*; функция $f(X)$ — *функцией регрессии*. Относительно случайной величины e обычно делается предположение, что она имеет нормальное распределение с нулевым математическим ожиданием.

3.4.1. Выбор функции регрессии

Выбор наилучшей в некотором смысле функции $f(X)$ составляет задачу регрессионного анализа. Но сначала необходимо установить критерий, с помощью которого можно определить, что такое "наилучшая" функция регрессии.

Одним из широко применяемых на практике критериев оптимальности функции регрессии является критерий *минимума суммы квадратов*. Он формулируется

следующим образом. Пусть имеются наблюдения $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Функция $f(x)$ подбирается таким образом, чтобы сумма квадратов

$$(y_1 - f(x_1))^2 + (y_2 - f(x_2))^2 + \dots + (y_n - f(x_n))^2$$

была минимальной. При определении функции регрессии этот критерий позволяет использовать хорошо разработанный *метод наименьших квадратов*, обеспечивающий построение функции регрессии, характеризуемой минимальным средним квадратом ее отклонения от экспериментальных данных.

Определив критерий оптимальности регрессии, следует перейти к выбору типа функции регрессии. Тип функции регрессии в значительной мере зависит от экспериментальных данных, однако наиболее часто используют многочлен вида $Y = a + b_1X + b_2X^2 + \dots + b_mX^m$ (коэффициенты a и b_i определяются на основе экспериментальных данных). Такая функция регрессии называется *полиномиальной*.

Остановимся на проблеме выбора степени многочлена. Выбор оптимальной степени аппроксимирующего многочлена зависит от многих факторов. Во-первых, от свойств аппроксимируемой функции $Y = f(X)$ (от ее гладкости [11]); во-вторых, от статистических характеристик наблюдаемых значений Y (особенно от дисперсии). Если априорная информация о функции $f(X)$ и статистических характеристиках наблюдаемых значений минимальна или отсутствует, то на практике считается, что степень многочлена не менее чем на порядок должна быть меньше числа точек данных, но не более 6-8. Обычно используют многочлены небольшой степени, часто — первой или второй.

Также часто применяются функции вида

- $Y = a + b \ln(X)$;
- $Y = a + bX + c \frac{1}{X}$;
- $Y = \frac{1}{a + bX}$ или $\frac{1}{Y} = a + bX$;
- $Y = \frac{1}{a + b_1X + b_2X^2 + \dots + b_mX^m}$ или $\frac{1}{Y} = a + b_1X + b_2X^2 + \dots + b_mX^m$;
- $Y = e^{a + bX}$ или $\ln(Y) = a + bX$;
- $Y = aX^b$ или $\ln(Y) = a_1 + b_1 \ln(X)$.

Обращаем внимание, что все приведенные функции или их преобразования линейны относительно коэффициентов. В общем случае такие функции можно представить в виде

$$\Psi(Y) = b_0\varphi_0(X) + b_1\varphi_1(X) + b_2\varphi_2(X) + \dots + b_m\varphi_m(X).$$

Здесь функции Ψ и φ заданы и, как правило, обладают "хорошими" свойствами, например дифференцируемостью. Коэффициенты b_i определяются на основе экспериментальных данных. Линейность относительно коэффициентов b_i данных функций значительно упрощает вычисление значений этих коэффициентов.

Конечно, при необходимости можно использовать функции, нелинейные относительно неизвестных параметров (коэффициентов). Они называются функциями *нелинейной регрессии*. В этом случае критерий минимума суммы квадратов также сохраняет свою силу, но непосредственное вычисление значений этих

неизвестных параметров резко усложняется — необходимо применять методы нелинейной оптимизации. Кроме того, возрастает сложность исследования статистических характеристик вычисленных параметров и уравнения регрессии в целом.

Если переменная X является вектором, т.е. $X = (X_1, X_2, \dots, X_n)$, то имеем так называемую *множественную регрессию*: функция регрессии здесь может зависеть как от отдельных компонентов вектора X , так и от любой комбинации этих компонентов. Простейшими функциями множественной регрессии являются полиномы вида

$$Y = a + \sum_{i=1}^n b_i X_i + \sum_{i=1}^n c_i X_i^2 + \dots + \sum_{i=1}^n d_i X_i^m + \sum_{i,j,\dots,k} e_{ij\dots k} X_i^{m_i} X_j^{m_j} \dots X_k^{m_k}$$

Здесь последняя сумма представляет всевозможные произведения переменных X_1, X_2, \dots, X_n , в разных степенях. Наибольшая степень переменных X , или суммы степеней их произведений называется степенью полинома. Отметим, что здесь функция регрессии также линейна относительно коэффициентов полинома. На практике редко используются такого типа полиномы степени, большей 2 или 3.

После выбора типа функции регрессии необходимо вычислить параметры этой функции и проверить адекватность построенной функции имеющимся данным, на основе которых рассчитывались параметры. Этим вопросам посвящены следующие разделы. В рамках регрессионного анализа также решаются задачи проверки значимости регрессии, построения доверительных интервалов для коэффициентов функции регрессии и проверки гипотез о значениях этих коэффициентов, вычисления значения переменной Y при тех значениях X , которых нет в исходных данных (задача прогнозирования) и др. Перечисленные задачи будут кратко рассмотрены ниже. Практические методы их решения описаны в главе 15.

3.4.2. Построение функции регрессии

Рассмотрим метод наименьших квадратов в самом общем случае, когда используется аппроксимация вида

$$\Psi(Y) = b_0 \varphi_0(X) + b_1 \varphi_1(X) + b_2 \varphi_2(X) + \dots + b_m \varphi_m(X).$$

Критерий минимума суммы квадратов запишется как

$$\sum_{i=1}^n (\Psi(y_i) - b_0 \varphi_0(x_i) - b_1 \varphi_1(x_i) - \dots - b_m \varphi_m(x_i))^2 = \min.$$

Здесь $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ — исходные данные. Для определения неизвестных коэффициентов b_0, b_1, \dots, b_m последнее выражение следует продифференцировать по этим коэффициентам и полученные производные приравнять к нулю. Получим так называемую *систему нормальных уравнений*:

$$\begin{cases} b_0 \sum_{i=1}^n \varphi_0^2(x_i) + b_1 \sum_{i=1}^n \varphi_0(x_i) \varphi_1(x_i) + \dots + b_m \sum_{i=1}^n \varphi_0(x_i) \varphi_m(x_i) = \sum_{i=1}^n \Psi(y_i) \varphi_0(x_i); \\ b_0 \sum_{i=1}^n \varphi_0(x_i) \varphi_1(x_i) + b_1 \sum_{i=1}^n \varphi_1^2(x_i) + \dots + b_m \sum_{i=1}^n \varphi_1(x_i) \varphi_m(x_i) = \sum_{i=1}^n \Psi(y_i) \varphi_1(x_i); \\ \dots \\ b_0 \sum_{i=1}^n \varphi_0(x_i) \varphi_m(x_i) + b_1 \sum_{i=1}^n \varphi_1(x_i) \varphi_m(x_i) + \dots + b_m \sum_{i=1}^n \varphi_m^2(x_i) = \sum_{i=1}^n \Psi(y_i) \varphi_m(x_i). \end{cases}$$

Значения коэффициентов b_0, b_1, \dots, b_m определяются как решения этой системы линейных алгебраических уравнений. За исключением редких случаев

вырожденности системы нахождение решения не представляет особых трудностей — в Excel имеется несколько средств решения таких систем (см. главу 6).

При простейшей аппроксимации многочленом первой степени уравнение линейкой *регрессии* имеет вид $Y = a + bX$. В нем следует определить значения коэффициентов a и b , удовлетворяющие критерию минимума суммы квадратов. В данном случае этот критерий запишется как

$$J^2(j, -a - b \cdot x_i) = \min.$$

Нормальная система уравнений будет иметь вид

$$\begin{cases} an + b \sum_{i=1}^n x_i = \sum_{i=1}^n y_i; \\ a \sum_{i=1}^n x_i + b \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i \end{cases}$$

Решением этой системы будут следующие формулы для вычисления коэффициентов a и b :

$$a = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n x_i y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}, \quad b = \frac{\sum_{i=1}^n y_i \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n x_i y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}.$$

При аппроксимации исходных данных многочленами более высоких степеней применяется подобный способ вычисления неизвестных коэффициентов многочлена. В Excel имеется несколько средств, позволяющих вычислять коэффициенты как линейной регрессии, так и полиномиальной. Эти средства описаны в части II, а их практическое применение — в главе 15.

3.4.3. Проверка адекватности функции регрессии

Пусть на основе экспериментальных данных $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ построена функция регрессии $f(x)$, зависящая от k параметров, значения которых рассчитываются по исходным данным. Обозначим через \hat{y}_i значения функции $f(x)$ в точках x_1, x_2, \dots, x_n : $\hat{y}_i = f(x_i)$, $i = 1, 2, \dots, n$. Для проверки адекватности функции регрессии исходным данным вычисляется дисперсионная таблица следующего вида.

Источник вариации (компоненты дисперсии)	Сумма квадратов	Число степеней свободы	Дисперсия
Регрессия	$SS_1 = \sum_{i=1}^n (\hat{y}_i - \bar{y})^2$	k	$s_f^2 = \frac{SS_1}{n}$
Остатки	$SS_2 = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$n - k - 1$	$s_e^2 = \frac{SS_2}{n - k - 1}$
Полная (общая) вариация	$SS = \sum_{i=1}^n (y_i - \bar{y})^2$ $SS = \sum_{i=1}^n y_i^2 - n \bar{y}^2$	$n - 1$	$s_Y^2 = \frac{SS}{n}$

Здесь $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$. Доказано, что $s_y^2 = s_f^2 + s_e^2$. Мерой адекватности функции

регрессии имеющимся данным служит величина $R^2 = \frac{s_f^2}{s_y^2}$, которая называется

коэффициентом детерминации. Этот коэффициент принимает значения от 0 до 1 и показывает, насколько велико общее отклонение значений функции регрессии от фактических значений величины Y . Если найдена идеальная функция регрессии, то $R^2 = 1$ (максимальное значение). В случае линейной регрессии R^2 равно квадрату коэффициента корреляции между случайными величинами X и Y ; корень из R^2 называется *индексом корреляции* I_{YX} (см. раздел 3.3.1). Таким образом, чем ближе коэффициент детерминации к 1, тем более точно выбранная функция регрессии соответствует экспериментальным данным.

Если случайная величина ϵ из уравнения зависимости $Y(X) = f(X) + \epsilon$ имеет нормальное распределение с нулевым математическим ожиданием, то существует критерий проверки значимости коэффициента детерминации. В этом случае

и при условии справедливости нулевой гипотезы $H_0: R^2 = 0$ статистика $F = \frac{s_f^2}{s_e^2}$

имеет F -распределение со степенями свободы k и $(n - k - 1)$. (Статистика F свя-

зана с коэффициентом R^2 соотношением $F = \frac{n-k-1}{k} \cdot \frac{R^2}{1-R^2}$.) Если найдена кван-

тиль t порядка $1 - \alpha$ (α — заданный уровень значимости) F -распределения со степенями свободы k и $(n - k - 1)$, то нулевая гипотеза принимается, если $F \leq t$. В противном случае принимается гипотеза о статистической значимости регрессии.

Кроме коэффициента детерминации, используются другие показатели адекватности функции регрессии исходным данным, в частности упоминаемый выше индекс корреляции. Также часто используется показатель *средней относительной ошибки аппроксимации*

$$\bar{\epsilon} = \frac{1}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right|.$$

lit Y ,

Чем меньше этот показатель, тем лучше функция регрессии аппроксимирует экспериментальные данные.

3.4.4. Статистические характеристики параметров функции регрессии

В регрессионном анализе относительно коэффициентов функции регрессии решаются следующие задачи.

1. Проверка значимости каждого коэффициента регрессии. Если значения коэффициентов регрессии статистически незначимы, то их следует исключить из уравнения регрессии.
2. Построение доверительных интервалов для значимых коэффициентов регрессии. Доверительные интервалы показывают точность вычисленных значений коэффициентов.

Эти задачи обычно решаются в предположении, что случайная ошибка ε в уравнении регрессионной зависимости имеет нормальное распределение с нулевым математическим ожиданием, а случайные погрешности ε_i каждого измерения y_i (реализации случайной величины ε) независимы и имеют одинаковые дисперсии. Если эти предположения выполняются, то вычисленные коэффициенты являются несмещенными и состоятельными² оценками истинных коэффициентов и асимптотически имеют нормальные распределения. В данном случае для проверки их значимости и построения доверительных интервалов используются стандартные методы, основанные на распределении Стьюдента. Эти методы описаны в разделе 15.3.

Отметим, что для решения обеих задач используются оценки дисперсий вычисленных коэффициентов функции регрессии. Формулы для определения этих оценок известны и внешне несложны, если использовать матричные обозначения. Приведем формулы для дисперсий коэффициентов a и b уравнения линейной регрессии $Y = a + bX$:

$$Da = \frac{\sigma^2 \sum_{i=1}^n x_i^2}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}, \quad Db = \frac{\sigma^2 n}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}.$$

Здесь σ^2 — дисперсия случайной величины ε (напомним зависимость $Y(X) = f(X) + \varepsilon$). При вычислениях σ^2 заменяют величиной s_ε^2 из дисперсионной таблицы.

Для вычисления дисперсий коэффициентов полиномиальной регрессии в Excel есть специальные средства, которые будут представлены в главе 4 (функция ЛИНЕЙН) и главе 5 (средство Регрессия) и использованы в главе 15.

3.4.5. Прогнозирование

Регрессионный анализ часто применяется для определения значения переменной Y в некоторой точке x_0 , не входящей в исходное множество значений $\{x_1, x_2, \dots, x_n\}$ переменной X . Для этого используется построенная функция регрессии $f(X)$ и значением переменной Y в точке x_0 считается величина $\hat{y} = f(x_0)$.

С точки зрения математика, здесь необходимо различать две возможные ситуации.

- Точка x_0 принадлежит интервалу, ограниченному минимальным и максимальными значениями множества $\{x_1, x_2, \dots, x_n\}$, если переменная X одномерна. В случае, когда переменная X является вектором, многомерная точка x_0 принадлежит выпуклой области, также определенной исходными значениями переменной X . В этой ситуации задача определения значения переменной Y называется *задачей восстановления значений* и является вполне корректной с математической точки зрения.
- Точка x_0 не принадлежит интервалу, определенному минимальным и максимальными значениями множества $\{x_1, x_2, \dots, x_n\}$ (переменная X одномерна), или соответствующей области для многомерной переменной X . В этой

² Небольшое уточнение: для состоятельности оценок дополнительно необходимо условие невырожденности матрицы, составленной из всевозможных попарных произведений $X_i X_j$.

ситуации задача определения значения переменной Y называется *задачей экстраполяции* (или *задачей прогнозирования*; этот термин часто используется при экономической интерпретации исходных данных) и в общем случае является некорректной с математической точки зрения.

В статистике обычно эти две ситуации четко не различаются, но их необходимо учитывать при проведении практического анализа. В общем случае несмотря на идеальную подгонку функции регрессии к исходным данным решение задачи прогнозирования может быть как угодно далеким от истинного значения $Y(x_0)$, если не накладывать априорных предположений о гладкости функции, описывающей истинную зависимость между X и Y . Доверительные интервалы, которые обычно строятся для величины $\hat{y} = f(x_0)$ (см. раздел 15.4), строго говоря, имеют право на существование только для задачи восстановления значений, хотя их применяют в обеих ситуациях. На практике прогнозирование можно применять достаточно “безопасно”, если зависимость между X и Y можно описать гладкой функцией, хотя бы дифференцируемой, и точка x_0 расположена недалеко от области, определяемой имеющимися значениями переменной X . (Но здесь возникнет вопрос, что понимать под словом *недалеко*.)

3.5. Дисперсионный анализ

Дисперсионный анализ³ — это статистический метод анализа результатов наблюдений, зависящих от различных факторов, определение наиболее влияющих факторов и оценка этого влияния. Факторами обычно называют внешние условия, влияющие на результаты наблюдений. Дисперсионный анализ заключается в разложении общей вариации (дисперсии) наблюдаемой случайной величины на отдельные складываемые, каждое из которых характеризует влияние того или иного фактора.

3.5.1. Статистическая модель

Дисперсионный анализ применяется в условиях следующей статистической модели. Наблюдаются n случайных величин X_1, X_2, \dots, X_n , каждая из которых представима в виде $X_i = \mu + \beta_1 + \beta_2 + \dots + \beta_m + \varepsilon_i$, где μ — константа (общее среднее), β_j — значение j -го фактора, ε_i — “остаточная” случайная величина, представляющая ошибки наблюдений, влияние неучтенных факторов и т.п. Как правило, предполагается, что случайные величины ε_i независимы между собой, одинаково распределены по нормальному закону с нулевым математическим ожиданием. Факторы обычно являются классификационными или порядковыми величинами, принимающими конечное множество значений. В таком случае, когда β_j принимает конкретное k -е значение из этого множества, говорят о k -м уровне j -го фактора.

Цель дисперсионного анализа заключается в оценке адекватности модели имеющимся выборочным значениям (для чего определяются статистические характеристики случайных величин ε_i), а также в оценке влияния факторов (другими словами, проверяются гипотезы о равенстве математических ожиданий случайных величин X_1, X_2, \dots, X_n).

В современной русской статистической литературе дисперсионный анализ иногда называют методом ANOVA (от англ. Analysis of Variance — анализ дисперсий).

Модель, в которой все β_j являются детерминированными, называется *моделью с постоянными факторами*. Если все β_j являются случайными величинами, такая модель называется *моделью со случайными факторами*. В случае, когда среди β_j есть как детерминированные, так и случайные величины, говорят о *смешанной модели*. В зависимости от количества факторов различают *однофакторный* и *многофакторный* (двухфакторный, трехфакторный и т.д.) дисперсионный анализ.

Прежде чем применять методы дисперсионного анализа, необходимо проверить обязательное условие: дисперсии всех исследуемых выборок должны быть одинаковыми. Если имеются только две выборки, для этого можно применить критерий Фишера сравнения дисперсий (описание критерия приводится в разделе 2.4.2, а его практическая реализация — в разделе 12.3.2). Для проверки равенства дисперсий нескольких выборок используется критерий Бартлетта (см. раздел 12.3.2).

3.5.2. Однофакторный дисперсионный анализ

Пусть имеется k выборок $(x_{11}, x_{12}, \dots, x_{1n_1}), (x_{21}, x_{22}, \dots, x_{2n_2}), \dots, (x_{k1}, x_{k2}, \dots, x_{kn_k})$ объемом соответственно n_1, n_2, \dots, n_k , которые являются реализациями случайных величин X_1, X_2, \dots, X_k . Предполагается, что каждая случайная величина X_i представима в виде $X_i = \mu + \beta + \varepsilon_i$, где β — фактор, который может принимать конечное множество значений (каждое значение фактора называется уровнем фактора), ε_i — “остаточная” случайная величина, имеющая нормальное распределение с нулевым математическим ожиданием. Все случайные величины ε_i независимы.

Считается, что каждая выборка соответствует одному уровню фактора. Для каждой выборки по стандартным формулам (см. раздел 2.1) подсчитывается выборочное среднее $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k$, выборочные дисперсии $S_1^2, S_2^2, \dots, S_k^2$ и об-

щее среднее $\bar{x} = \frac{1}{n} \sum_{i=1}^k \bar{x}_i$, где $n = \sum_{i=1}^k n_i$.

Далее рассчитываются суммы квадратов:

$$\text{межгрупповая сумма квадратов } SS_1 = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2,$$

$$\text{внутригрупповая сумма квадратов } SS_2 = \sum_{i=1}^k (n_i - 1) S_i^2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2,$$

$$\text{полная сумма квадратов } SS = SS_1 + SS_2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2.$$

Межгрупповую сумму квадратов также называют *рассеиванием по факторам*, внутригрупповую сумму квадратов — *остаточным рассеиванием*, полную сумму квадратов — *полной (или общей) суммой квадратов отклонений отдельных наблюдений от общего среднего \bar{x}* . Для каждой из этих сумм определяются *степени свободы*: для межгрупповой суммы квадратов число степеней свободы равно $k - 1$, для внутригрупповой суммы квадратов — $n - k$, а для полной суммы — $n - 1$. На основании значений сумм квадратов и соответствующих степеней свободы вычисляются следующие дисперсии:

$$\text{межгрупповая дисперсия } s_1^2 = \frac{SS_1}{k-1} = \frac{1}{k-1} \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2,$$

$$\text{внутригрупповая дисперсия } s_2^2 = \frac{SS_2}{n-k} = \frac{1}{n-k} \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2,$$

$$\text{полная дисперсия } s^2 = \frac{SS}{n-1} = \frac{1}{n-1} \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2.$$

$$\text{дисперсия } s^2 = \frac{Y < YJ(XU^{-X})^2}{-} +$$

Межгрупповая дисперсия показывает, насколько различаются выборочные средние. Она равна нулю, если средние равны, и чем сильнее различаются средние в разных выборках, тем она больше. Эта дисперсия является мерой разброса выборочных средних вследствие влияния фактора. Внутригрупповая дисперсия показывает, насколько неоднородна каждая выборка (группа). Она показывает влияние неучтенных "остаточных" факторов (величин ϵ).

Проведенные вычисления принято представлять в виде *таблицы дисперсионного анализа*, или в *дисперсионной таблице*, следующего вида.

<i>Источник вариации (компоненты дисперсии)</i>	<i>Сумма квадратов</i>	<i>Число степеней свободы</i>	<i>Дисперсия</i>
Межгрупповая вариация (различия между выборками)	$SS_1 = \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2$	$ft \quad k - 1$	$s_1^2 = \frac{SS_1}{k - 1}$
Внутригрупповая вариация (различия внутри выборок)	$SS_2 = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2$	$n - k$	$s_2^2 = \frac{SS_2}{n - k}$
Полная (общая) вариация	$SS = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x})^2$	$n - 1$	$s^2 = \frac{SS}{n - 1}$

Отношение $R^2 = s_1^2 / s^2$ называется *коэффициентом детерминации* и показывает, какая часть полной дисперсии объясняется влиянием фактора.

Для проверки гипотезы о равенстве математических ожиданий во всех выборках применяют критериальную статистику $F = s_1^2 / s_2^2$, которая в случае истинности проверяемой гипотезы имеет F -распределение со степенями свободы $(k - 1)$ и $(n - k)$. При заданном доверительном уровне α критическое значение $t_{кр}$ определяется как квантиль порядка $1 - \alpha$ этого распределения. Если $F < t_{кр}$, то гипотеза о равенстве математических ожиданий не отклоняется (другими словами, влияние фактора во всех выборках одинаково). В случае $F \geq t_{кр}$ эта гипотеза отвергается. Отметим, что статистика $F \geq 1$, в противном случае гипотеза о равенстве математических ожиданий принимается без проверки, поскольку тогда различия между выборками меньше различий внутри выборок, т.е. влияние фактора менее значимо, чем влияние "остаточных" факторов.

Если гипотеза о равенстве математических ожиданий отвергается, необходимо узнать, математические ожидания каких выборок значимо отличаются от других. В этом случае нельзя непосредственно попарно сравнивать отдельные выборки с помощью, например, критерия Стьюдента (см. раздел 2.4.2), поскольку резко возрастает *групповая ошибка* первого рода (т.е. возрастает вероятность того, что по крайней мере один из тестов неверно отвергнет нулевую гипотезу). В такой ситуации для попарных сравнений следует применять либо метод множественных сравнений Шеффе (описанный в разделе 12.3.1), либо модифицированный критерий Стьюдента.

Модификация критерия Стьюдента заключается в том, что критериальная статистика вычисляется по формуле $T = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{SS_2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}$, где \bar{x}_1 и \bar{x}_2 — средние

сравниваемых выборок, n_1 и n_2 — объемы этих выборок, а SS_2 — внутригрупповая сумма квадратов, которая рассчитывается в дисперсионной таблице по всем выборкам. Для определения критического значения берется квантиль порядка $1 - \alpha/2$ распределения Стьюдента с $(n - k)$ степенью свободы (обратите внимание на значение степени свободы).

3.5.3. Двухфакторный дисперсионный анализ

В двухфакторном дисперсионном анализе предполагается, что случайные величины X_i представимы в виде $X_i = \mu + \beta + \gamma + \varepsilon_i$, где μ — константа (общее среднее), β и γ — значения факторов, ε_i — “остаточная” случайная величина, имеющая стандартные статистические характеристики (т.е. все ε_i независимы и имеют одинаковые нормальные распределения с нулевым математическим ожиданием и одинаковыми дисперсиями).

При использовании двухфакторного анализа возможны две ситуации. В первой ситуации имеется одна выборка, в которой каждое выборочное значение соответствует одной комбинации уровней факторов β и γ , во второй ситуации имеется несколько подобных выборок⁴. Рассмотрим сначала применение дисперсионного анализа в условиях первой ситуации.

Двухфакторный дисперсионный анализ без повторений

Итак, имеется двумерная выборка, состоящая из выборочных значений x_{ij} , индекс i соответствует i -му уровню фактора β (будем обозначать уровни фактора β как β_i), индекс j соответствует j -му уровню фактора γ (уровни этого фактора обозначим как γ_j). Пусть фактор β имеет r уровней, а фактор γ — t уровней. Таким образом, в общем случае выборка имеет размерность $r \times t$ ⁵. Такую выборку удобно представлять в виде таблицы⁶.

⁴ Сразу скажем, что в условиях первой ситуации применяется средство Excel Двухфакторный дисперсионный анализ без повторений, а в условиях второй — Двухфакторный дисперсионный анализ с повторениями (см. главу 5). Эти две ситуации не совсем четко представлены в описании данных средств в справочной системе Excel.

⁵ При чтении материала о двухфакторном анализе полезно иметь в виду “классическую” модель. Имеется ряд сельскохозяйственных полей (участков), к каждому из них применяется свой способ обработки земли (есть r различных способов обработки) и вносится t различных удобрений (на каждый участок вносится один тип удобрения). Требуется исследовать урожайность некоторой сельскохозяйственной культуры в зависимости от двух факторов — способа обработки земли и вида удобрения. В данном случае значением случайной величины X является урожайность на каждом опытном участке. “Классика” здесь заключается в том, что дисперсионный анализ первоначально был предложен Р. Фишером (Fisher, 1925) для обработки результатов агрономических опытов по определению условий, при которых испытываемый сорт сельскохозяйственной культуры дает наибольший урожай.

⁶ Эта таблица напоминает таблицу сопряженности (см. раздел 3.3.3), но величины x_{ij} в них имеют разный смысл — в таблице сопряженности величина x_{ij} равна количеству наблюдений, соответствующих признакам A_i и B_j , здесь же x_{ij} — значение случайной величины X .

	γ_1	γ_2	...	γ_t	Средние
β_1	x_{11}	x_{12}	...	x_{1t}	$\bar{x}_{1.} = \frac{1}{t} \sum_{i=1}^t x_{1i}$
β_2	x_{21}	x_{22}	...	x_{2t}	$\bar{x}_{2.} = \frac{1}{t} \sum_{i=1}^t x_{2i}$
...
β_r	x_{r1}	x_{r2}	...	x_{rt}	$\bar{x}_{r.} = \frac{1}{t} \sum_{i=1}^t x_{ri}$
Средние	$\bar{x}_{.1} = \frac{1}{r} \sum_{j=1}^r x_{j1}$	$\bar{x}_{.2} = \frac{1}{r} \sum_{j=1}^r x_{j2}$...	$\bar{x}_{.t} = \frac{1}{r} \sum_{j=1}^r x_{jt}$	$\bar{x} = \frac{1}{rt} \sum_{i=1}^r \sum_{j=1}^t x_{ij}$

Точечной оценкой общего среднего μ является величина $\bar{x} = \frac{1}{rt} \sum_{i=1}^r \sum_{j=1}^t x_{ij}$. Величины $\bar{x}_{k.} = \frac{1}{t} \sum_{i=1}^t x_{ki}$ и $\bar{x}_{.m} = \frac{1}{r} \sum_{j=1}^r x_{jm}$ называются средними по уровням факторов: $\bar{x}_{k.}$ — среднее по уровню k фактора β , $\bar{x}_{.m}$ — среднее по уровню m фактора γ .

Определение степени влияния факторов выполняется так же, как и в однофакторном анализе, на основе дисперсионной таблицы, которая для двухфакторного анализа без повторений имеет следующую структуру.

Источник вариации (компоненты дисперсии)	Сумма квадратов	Число степеней свободы	Дисперсия
Вариация между средними по строкам (различия между уровнями фактора β)	$SS_1 = t \sum_{i=1}^r (\bar{x}_{i.} - \bar{x})^2$	$r - 1$	$s_1^2 = \frac{SS_1}{r - 1}$
Вариация между средними по столбцам (различия между уровнями фактора γ)	$SS_2 = r \sum_{j=1}^t (\bar{x}_{.j} - \bar{x})^2$	$t - 1$	$s_2^2 = \frac{SS_2}{t - 1}$
Остаточная вариация (различия внутри выборки)	$SS_3 = \sum_{i=1}^r \sum_{j=1}^t (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2$	$(r - 1)(t - 1)$	$s_3^2 = \frac{SS_3}{(r - 1)(t - 1)}$
Полная (общая) вариация	$SS = \sum_{i=1}^r \sum_{j=1}^t (x_{ij} - \bar{x})^2$	$rt - 1$	$s^2 = \frac{SS}{rt - 1}$

Для определения степени влияния факторов сравнивают дисперсии по факторам с остаточной дисперсией. Например, для проверки нулевой гипотезы о равенстве средних по уровням фактора β (если эта гипотеза принимается, то влияние фактора β незначимо) вычисляются сначала критериальная статистика $T = s_1^2/s_3^2$, а затем —

критическое значение $t_{кр}$ как квантиль порядка $1 - \alpha$ (α — заданный уровень значимости) F -распределения со степенями свободы $(r - 1)$ и $(r - 1)(t - 1)$. Если выполняется неравенство $T < t_{кр}$, то нулевая гипотеза принимается, в противном случае — отклоняется. Аналогично определяется степень влияния фактора γ .

В Excel двухфакторный дисперсионный анализ без повторений выполняет одноименное средство из пакета анализа (см. раздел 5.13). Выполнение этого анализа без применения данного средства показано в разделе 14.3.1.

Двухфакторный дисперсионный анализ с повторениями

Эта разновидность дисперсионного анализа предполагает, что имеется несколько двумерных выборок такого же вида, что и в двухфакторном анализе без повторений. Здесь все выборочные значения также можно представить в виде таблицы, как и в случае одной выборки. Но теперь в каждой ячейке этой таблицы, соответствующей i -му уровню фактора β и j -му уровню фактора γ , будет находиться не одно значение x_{ij} , а m значений x_{ijk} ($k = 1, \dots, m$), m — количество выборок. Предполагаем, что объемы всех выборок одинаковы, т.е. в каждой ячейке таблицы содержится одинаковое количество значений. Если это не так, то приведенные ниже формулы несколько усложняются [24].

По выборочным значениям вычисляются:

- средние по каждой ячейке $\bar{x}_{ij\cdot} = \frac{1}{m} \sum_{k=1}^m x_{ijk}$;
- средние по строкам $\bar{x}_{i\cdot\cdot} = \frac{1}{t} \sum_{j=1}^t \bar{x}_{ij\cdot}$;
- средние по столбцам $\bar{x}_{\cdot j\cdot} = \frac{1}{r} \sum_{i=1}^r \bar{x}_{ij\cdot}$;
- общее среднее $\bar{x} = \frac{1}{rt} \sum_{i=1}^r \sum_{j=1}^t \bar{x}_{ij\cdot}$.

Порядок проведения дисперсионного анализа в данном случае такой же, как и прежде: сначала вычисляются суммы квадратов, затем вычисляются оценки дисперсий, далее для проверки гипотез о влиянии факторов вычисляются отношения дисперсий, которые сравниваются с критическими значениями, полученными как квантили F -распределения. Дисперсионная таблица имеет следующий вид.

Источник вариации (компоненты дисперсии)	Сумма квадратов	Число степеней свободы	Дисперсия
Вариация между средними по строкам (различия между уровнями фактора β)	$SS_1 = mt \sum_{i=1}^r (\bar{x}_{i\cdot\cdot} - \bar{x})^2$	$r - 1$	$s_1^2 = \frac{SS_1}{r - 1}$
Вариация между средними по столбцам (различия между уровнями фактора γ)	$SS_2 = mr \sum_{j=1}^t (\bar{x}_{\cdot j\cdot} - \bar{x})^2$	$t - 1$	$s_2^2 = \frac{SS_2}{t - 1}$

Источник вариации (компоненты дисперсии)	Сумма квадратов	Число степеней свободы	Дисперсия
Взаимодействие факторов β и γ	$SS_3 = m \sum_{i=1}^r \sum_{j=1}^t (\bar{x}_{ij\cdot} - \bar{x}_{i\cdot\cdot} - \bar{x}_{\cdot j\cdot} + \bar{x})^2$	$(r-1)(t-1)$	$s_3^2 = \frac{SS_3}{(r-1)(t-1)}$
Остаточная вариация	$SS_4 = \sum_{k=1}^m \sum_{i=1}^r \sum_{j=1}^t (x_{ijk} - \bar{x}_{ij\cdot})^2$	$rt(m-1)$	$s_4^2 = \frac{SS_4}{rt(m-1)}$
Полная (общая) вариация	$SS = \sum_{k=1}^m \sum_{i=1}^r \sum_{j=1}^t (x_{ijk} - \bar{x})^2$	$rtm-1$	$s^2 = \frac{SS}{rtm-1}$

Как видно из этой таблицы, появился новый источник вариации, а именно — эффект от взаимодействия факторов β и γ . Этот эффект не наблюдается в случае одной выборки, поскольку там разности $x_{ijk} - \bar{x}_{ij\cdot}$ равны нулю. Обычно неявно предполагается, что это взаимодействие можно описать как произведение $\delta\beta_i\gamma_j$ и дисперсионный анализ должен определить, насколько значимо величина δ отличается от нуля. Необходимо отметить, что существуют и другие модели взаимодействия факторов [1].

В Excel двухфакторный дисперсионный анализ с повторениями выполняет одноименное средство из пакета анализа (см. раздел 5.12).

Многофакторный дисперсионный анализ выполняется аналогично двухфакторному: вычисляются сначала средние по каждому уровню всех факторов, затем — суммы квадратов для каждого фактора и взаимодействий всех возможных комбинаций факторов, после — степени свободы и соответствующие дисперсии. Далее для проверки гипотез о влиянии факторов вычисляются отношения дисперсий факторов или их взаимодействий к остаточной дисперсии (это критериальные статистики) и находятся критические значения как квантили F -распределения с соответствующими степенями свободы. Если значения критериальных статистик меньше критических значений, то нулевые гипотезы принимаются, в противном случае — отвергаются. Подробно многофакторный дисперсионный анализ описан в [24].

Средства Excel для статистического анализа

В этой части...

Глава 4. Статистические функции

Глава 5. Надстройка Пакет анализа

Глава 6. Дополнительные возможности Excel для проведения статистического анализа

Глава 7. Моделирование случайных величин

В этой части описаны возможности Excel для проведения статистического анализа. Предполагается, что читатель знаком с основами работы в этой электронной таблице хотя бы в следующем объеме: он может вводить и редактировать данные, создавать формулы, использовать функции, строить диаграммы и графики, форматировать рабочий лист и т.п. Непосредственно для статистической обработки данных в Excel предусмотрены многочисленные статистические функции (около 80) и средства надстройки Пакет анализа. Статистические функции (глава 4) и средства, предоставляемые надстройкой Пакет анализа (глава 5), в данной части описаны достаточно полно. Кроме статистических функций и средств пакета анализа, здесь рассмотрены общие средства и надстройки Excel, которые также можно использовать в статистическом анализе (глава 6). Это формулы массивов, специального вида диаграммы и графики, а также надстройка Поиск решения. В главе 7 мы кратко остановимся на возможностях моделирования случайных величин в Excel.

Статистические функции

К статистическим функциям Excel обычно относят те функции, которые приведены в мастере функций (или в справочной системе Excel) в категории Статистические. Однако эта категория содержит также функции, которые скорее можно отнести к категории просто математических (например, функции МИН и МАКС) либо информационных функций (функции СЧЁТ и СЧЁТЗ). С другой стороны, в других категориях функций также имеются функции, которые можно использовать при проведении статистического анализа (например, некоторые функции для матричных вычислений). Поэтому мы разобьем категорию статистических функций на несколько групп функций, выполняющих однотипные действия (например, вычисляющие значения функций распределений либо выполняющие тесты), выделив группу дополнительных и вспомогательных функций. К сожалению, справочные разделы Excel, посвященные статистическим функциям, написаны весьма невнятно, имеют много неточностей, а порой просто содержат ошибки. Поэтому будем описывать эти функции по возможности полно.

4.1. Функции для определения экстремальных значений выборки

В эту группу функций входят следующие функции.

<i>Функция</i>	<i>Назначение</i>
МАКС	Возвращает максимальное значение из списка аргументов
МАКСА	Возвращает наибольшее значение из списка аргументов. Наряду с числовыми значениями выполняет также сравнение текстовых и логических значений
МИН	Возвращает минимальное значение из списка аргументов
МИНА	Возвращает наименьшее значение из списка аргументов. Наряду с числовыми значениями выполняет также сравнение текстовых и логических значений
НАИБОЛЬШИЙ	Возвращает k-е наибольшее значение из массива данных
НАИМЕНЬШИЙ	Возвращает k-е наименьшее значение из массива данных

4.1.1. Функции МАКС, МАКСА, МИН, МИНА

Эти функции имеют следующий синтаксис (далее, если будет приводиться синтаксис для *группы* функций, в описании синтаксиса будем использовать вместо конкретного названия функции слово **ФУНКЦИЯ**):

ФУНКЦИЯ(Число1;Число2;...)

Функции могут содержать до 30 аргументов. Аргументами могут быть конкретные числа, адреса диапазонов либо ссылки на отдельные ячейки рабочего листа. В диапазонах пустые ячейки и ячейки с текстом игнорируются. В функциях **МАКСА** и **МИНА** аргументы, содержащие значение **ИСТИНА**, интерпретируются как единица, а аргументы, содержащие значение **ЛОЖЬ** или текст, интерпретируются как нуль; в функциях **МАКС** и **МИН** такие аргументы игнорируются.

4.1.2. Функции НАИБОЛЬШИЙ и НАИМЕНЬШИЙ

Синтаксис функций:

ФУНКЦИЯ(Массив;к)

Аргумент **Массив** — ссылка на диапазон ячеек, из которого выбирается *k*-е наибольшее (наименьшее) числовое значение. Целое число *k* задает позицию (начиная с наибольшей в функции **НАИБОЛЬШИЙ** и с наименьшей в функции **НАИМЕНЬШИЙ**). Если аргумент **Массив** не задан либо если число *k* меньше 0 или больше количества ячеек в диапазоне **Массив**, то функции возвращают значение ошибки **#ЧИСЛО!**.

Покажем, как с помощью функций **НАИБОЛЬШИЙ** и **НАИМЕНЬШИЙ** определить выборочное значение, которому соответствует заданный ранг (ранг — номер позиции выборочного значения в вариационном ряде, построенном по выборке; о вариационном ряде и рангах выборочных значений речь идет в разделе 2.3.9). Отметим, что для решения данной задачи вариационный ряд заранее не строится — он получается в результате вычислений.

Пусть выборочные значения располагаются в столбце **A** (на рис. 4.1 выборочные значения получены с помощью функции **СЛЧИС**, умноженной на 10). В одном из соседних столбцов вводятся натуральные числа от 1 до *n* (*n* — объем выборки). В ячейку **D2** вводится формула **=НАИМЕНЬШИЙ(\$A\$2:\$A\$16;C2)**, которая затем копируется вниз. В результате получаем вариационный ряд, числа в столбце **C** показывают ранги значений этого ряда. Аналогичного результата можно добиться с помощью функции **НАИБОЛЬШИЙ**, для чего в ячейку вводится формула **=НАИБОЛЬШИЙ(\$A\$2:\$A\$16',16-C2)**, которая затем также копируется вниз (здесь число 16 — это число, на 1 большее объема выборки).

Отметим, что выбрать наибольшие или наименьшие значения из выборки (если выборочные значения расположены в одном столбце) можно также с помощью средства Excel **Автофильтр** (команда **Данные^Фильтр^Автофильтр**) или **Расширенный фильтр** (команда **Данные^Фильтр^Расширенный фильтр**), причем в этом случае можно выбрать не отдельные значения, а несколько значений, например 5 наибольших или 10 наименьших значений (можно количество выбираемых значений задать в виде процента от объема выборки). Недостатком использования фильтров является то, что это средство не "интерактивно"; поэтому, чтобы снова получить нужные значения при изменении некоторых выборочных значений, фильтр необходимо применить еще раз — в отличие от функций, которые автоматически пересчитываются при внесении изменений в выборку.

D2		=ПЕРСЕНТИЛЬ(\$A\$2:\$A\$15;C2)				
	A	B	C	D	E	F
1	Выборка		k	Функция ПЕРСЕНТИЛЬ		
2	9,9375		0	0,9223		
3	1,1390		0,1	1,0252		
4	8,3622		0,2	1,1865		
5	7,0857		0,3	1,6849		
6	1,2181		0,4	2,3539		
7	4,9355		0,5	3,1844		
8	3,2903		0,6	4,6064		
9	5,7185		0,7	5,6991		
10	3,0785		0,8	6,2654		
11	5,6969		0,9	7,9792		
12	1,7146		1	9,9375		
13	0,9765					
14	2,1728					
15	0,9223					
16						

Рис. 4.3. Функция ПЕРСЕНТИЛЬ

4.2.3. Функция МЕДИАНА

Эта функция возвращает медиану — квантиль порядка 0,5. Медиана в данном случае определяется как число, которое является серединой вариационного ряда $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$, т.е. половина чисел имеет значения, большие чем медиана, а половина — меньшие чем медиана. Если n — нечетное число ($n = 2k + 1$), то в качестве медианы берется число $x_{(k)}$; если же n — четное число ($n = 2k$), то медиана вычисляется как среднее чисел $x_{(k-1)}$ и $x_{(k)}$.

Синтаксис функции:

МЕДИАНА(Число1;Число2;...)

Функция может иметь 30 аргументов Число. Эти аргументы должны быть числами, массивами или ссылками на диапазоны ячеек, содержащих числа. Если в заданном диапазоне ячеек имеются ячейки, содержащие текст, логические значения или пустые ячейки, то они игнорируются; но ячейки, содержащие нулевые значения, учитываются.

4.2.4. Функция ПРОЦЕНТРАНГ

Данная функция вычисляет так называемый *процентный ранг* выборочных значений. Он вычисляется следующим образом. По выборке x_1, x_2, \dots, x_n строится вариационный ряд $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. Номер r члена ряда $x_{(r)}$ — это ранг значения $x_{(r)}$. Процентный ранг этого значения вычисляется по формуле $(r - 1)/(n - 1)$, где n — объем выборки. Можно считать, что данная функция обратная к функции ПЕРСЕНТИЛЬ, где по значению порядка квантиля (что практически совпадает со значением порядкового ранга) находится соответствующая процентиль, т.е. выборочное значение (если процентиль совпадает с выборочным значением).

Синтаксис функции:

ПРОЦЕНТРАНГ(Массив;х;Размерность)

Аргумент Массив — это числовой массив или адрес диапазона ячеек, содержащего выборочные значения. Аргумент x — значение, для которого вычисляется процентный ранг. Если это значение не совпадает с каким-либо выборочным значением, то функция ПРОЦЕНТРАНГ для этого значения вычисляет ранг как среднее рангов тех членов вариационного ряда, между которыми заключено данное значение x . Необязательный аргумент **Размерность** определяет количество десятичных знаков после запятой в вычисленном значении процентного ранга. Если этот аргумент опущен, то по умолчанию процентный ранг записывается с тремя десятичными знаками.

4.2.5. Функция РАНГ

Из самого названия функции РАНГ понятно, что данная функция вычисляет ранг выборочного значения, т.е. номер r значения $x_{(r)}$ вариационного ряда $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. Отметим, что функция может упорядочить выборку как по возрастанию, так и по убыванию (способ упорядочивания задает аргумент функции **Порядок**, о чем речь идет ниже), и, конечно, возвращаемые функцией значения будут различны для разных способов упорядочивания выборки. Если в выборке есть совпадающие значения, то им присваиваются одинаковые ранги, а последующему значению присваивается ранг, значение которого будет больше предыдущего ранга на количество одинаковых выборочных значений. Например, если в выборке дважды встречается число 10, имеющее ранг 5, то следующее по величине число 11 будет иметь ранг 7 и ни одно из чисел не будет иметь ранг 6 (пример взят из справочной системы Excel).

Синтаксис функции:

РАНГ(Число;Массив;Порядок)

Аргумент Массив — числовой массив или адрес диапазона ячеек, содержащего выборочные значения. Аргумент **Число** — значение, для которого вычисляется ранг. Если это значение не совпадает ни с одним выборочным значением, то функция возвращает значение ошибки #Н/Д!. Необязательный аргумент **Порядок** определяет способ упорядочивания выборки. Если **Порядок** равен 0 (нулю) или опущен, то выборка упорядочивается по убыванию. Если **Порядок** — любое ненулевое число, то выборка упорядочивается по возрастанию.

Отметим, что с помощью данной функции можно вычислить не только ранг одного выборочного значения, но одновременно ранги *всех* выборочных значений. Для этого надо применить ее в *формуле массива*. О формулах массива подробнее поговорим в главе 6, в разделе 6.1, здесь же просто покажем, как подсчитать все ранги одной выборки.

1. Пусть выборочные значения записаны в одном столбце А, как показано на рис. 4.4.
2. Вычисленные ранги будут записаны в столбце В. Выделите диапазон ячеек В2:В17 и в первую ячейку выделенного диапазона введите формулу $=\text{РАНГ}(A2:A17;A2:A17;1)$ (см. рис. 4.4).

3. Нажмите комбинацию клавиш <Ctrl+Shift+Enter> (ввод формулы массива). Ранги будут вычислены для всей выборки, как показано на рис. 4.5. Обратите внимание на равные ранги в ячейках В10, В11 и В13; выборочные значения в ячейках А10, А11 и А13 также одинаковы.

НАИБОЛЬШИЙ				
=РАНГ(A2:A17;A2:A17;1)				
1	Выборка	Ранги		
2	14	=РАНГ(A2:A17;A2:A17;1)		
3	12			
4	13			
5	7			
6	18			
7	8			
8	11			
9	5			
10	19			
11	19			
12	2			
13	19			
14	17			
15	6			
16	19			
17	12			

Рис. 4.4. Ввод формулы массива

B2				
=РАНГ(A2:A17;A2:A17;1)				
1	Выборка	Ранги		
2	14	10		
3	12	7		
4	13	9		
5	7	4		
6	18	12		
7	8	5		
8	11	6		
9	5	2		
10	19	13		
11	19	13		
12	2	1		
13	19	13		
14	17	11		
15	6	3		
16	19	13		
17	12	7		

Рис. 4.5. Вычисленные ранги

4.3. Функции для вычисления средних

Функции этой группы вычисляют средние значения: среднее арифметическое, среднее геометрическое и среднее гармоническое.

Функция	Назначение
СРГАРМ	Возвращает среднее гармоническое множества данных
СРГЕОМ	Возвращает среднее геометрическое множества данных
СРЗНАЧ	Возвращает среднее арифметическое своих аргументов
СРЗНАЧА	Вычисляет среднее арифметическое своих аргументов; помимо чисел, в расчете могут участвовать текстовые и логические значения
УРЕЗСРЕДНЕЕ	Возвращает среднее арифметическое, рассчитанное после отбрасывания крайних значений массива данных

Все перечисленные функции, кроме функции УРЕЗСРЕДНЕЕ, имеют следующий синтаксис:

ФУНКЦИЯ(Число1;Число2;...)

Они могут иметь до 30 аргументов Число. Этими аргументами могут быть или непосредственно числовые значения, или ссылки на диапазоны ячеек, содержащих значения, при этом пустые ячейки игнорируются, а ячейки с нулевыми значениями засчитываются. Функция СРЗНАЧА интерпретирует логическое значение ИСТИНА как 1, а логическое значение ЛОЖЬ и текстовые значения — как 0. Другие функции логические и текстовые значения игнорируют. Функции СРГАРМ и СРГЕОМ также требуют, чтобы все числовые значения, которые они обрабатывают, были положительными. Иначе они возвращают ошибку #ЧИСЛО!.

4.3.1. Функция СРГАРМ

Эта функция вычисляет среднее гармоническое H значений x_1, x_2, \dots, x_n по формуле $H \equiv \frac{n}{\sum_{i=1}^n \frac{1}{x_i}}$. Среднее гармоническое не превышает среднего геометрического, которое, в свою очередь, не превышает среднего арифметического.

4.3.2. Функция СРГЕОМ

Данная функция вычисляет среднее геометрическое G значений x_1, x_2, \dots, x_n по формуле $G = \sqrt[n]{x_1 \cdot x_2 \cdot \dots \cdot x_n}$.

4.3.3. Функции СРЗНАЧ и СРЗНАЧА

Эти функции вычисляют среднее арифметическое (выборочное среднее) \bar{x} значений x_1, x_2, \dots, x_n по формуле $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$. Различие между функциями СРЗНАЧ и СРЗНАЧА заключается только в интерпретации логических и текстовых

значений. Функция СРЗНАЧ их игнорирует, а СРЗНАЧА присваивает им числовые значения: значение 1 — логическому значению ИСТИНА и значение 0 — логическому значению ЛОЖЬ и текстовым значениям.

4.3.4. Функция УРЕЗСРЕДНЕЕ

Эта функция возвращает среднее арифметическое, рассчитанное после отбрасывания заданного количества крайних значений массива данных.

Синтаксис функции:

УРЕЗСРЕДНЕЕ(Массив;Доля)

Аргумент Массив — это числовой массив или адрес диапазона ячеек, содержащего данные. Аргумент Доля — это доля точек данных, исключаемых из вычислений, т.е. количество исключаемых точек вычисляется как Доля $\times n$, где n — общее количество точек данных. Данное произведение округляется с недостатком до ближайшего четного числа и половина этого числа представляет собой равные количества отбрасываемых наименьших и наибольших значений из массива данных. Если значение аргумента Доля отрицательно или больше 1, то функция возвращает значение ошибки #ЧИСЛО!.

4.4. Функции для вычисления геометрических характеристик распределения

В эту группу функций входят следующие функции.

Функция	Назначение
СКОС	Возвращает выборочный коэффициент асимметрии
ЭКССЕСС	Возвращает выборочный коэффициент эксцесса

Синтаксис функций:

ФУНКЦИЯ(Число1;Число2;...)

Они могут иметь до 30 аргументов Число. Этими аргументами могут быть непосредственно числовые значения, числовые массивы или ссылки на диапазоны ячеек, содержащих значения, при этом пустые ячейки, а также ячейки, содержащие логические и текстовые значения, игнорируются, но ячейки с нулевыми значениями засчитываются.

4.4.1. Функция СКОС

Эта функция вычисляет выборочный коэффициент асимметрии распределения (о коэффициенте асимметрии и его значении речь идет в разделе 1.2.3). Если есть выборка x_1, x_2, \dots, x_n (задается аргументами Число), функция СКОС вычисляет выборочный коэффициент асимметрии по следующей формуле:

$$\hat{\beta}_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_n} \right)^3,$$

где n — объем выборки, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$.

Если $n < 3$ или $s_n = 0$, то функция возвращает значение ошибки #ДЕЛ/0!.

4.4.2. Функция ЭКСЦЕСС

Эта функция вычисляет выборочный коэффициент эксцесса распределения (о коэффициенте эксцесса и его значении речь идет в разделе 1.2.3). Если есть выборка x_1, x_2, \dots, x_n (задается аргументами Число), то функция ЭКСЦЕСС вычисляет выборочный коэффициент эксцесса по следующей формуле:

$$\hat{\beta}_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_n} \right)^4 - \frac{3(n-1)^2}{(n-3)(n-3)},$$

где n — объем выборки, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, $s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$.

Если $n < 4$ или $s_n = 0$, то функция возвращает значение ошибки #ДЕЛ/0!.

4.5. Функции для вычисления выборочной дисперсии и отклонения

В эту группу включены функции, которые вычисляют ту или иную меру разброса выборочных значений относительно среднего.

Функция	Назначение
ДИСП	Вычисляет несмещенную оценку дисперсии выборки
ДИСПА	Вычисляет несмещенную оценку дисперсии выборки, учитывая логические и текстовые значения
ДИСПР	Вычисляет асимптотически несмещенную оценку дисперсии выборки
ДИСПРА	Вычисляет асимптотически несмещенную оценку дисперсии выборки, учитывая логические и текстовые значения
КВАДРОТКЛ	Возвращает сумму квадратов отклонений
СРОТКЛ	Возвращает среднее значение абсолютных величин отклонений точек данных от среднего
СТАНДОТКЛОН	Оценивает стандартное отклонение по выборке
СТАНДОТКЛОНА	Оценивает стандартное отклонение по выборке, в расчете также учитываются текстовые и логические значения
СТАНДОТКЛОНП	Вычисляет стандартное отклонение по генеральной совокупности
СТАНДОТКЛОНПА	Вычисляет стандартное отклонение по генеральной совокупности, в расчете также учитываются текстовые и логические значения

Синтаксис функций:

ФУНКЦИЯ(Число1;Число2;...)

Функции могут иметь до 30 аргументов Число. Этими аргументами могут быть непосредственно числовые значения, числовые массивы или ссылки на диапазоны ячеек, содержащих значения, при этом пустые ячейки игнорируются, а ячейки с нулевыми значениями засчитываются. Функции ДИСПА, ДИСПРА, СТАНДОТКЛОНА и СТАНДОТКЛОНПА интерпретируют логическое значение ИСТИНА как 1, а логическое значение ЛОЖЬ и текстовые значения — как 0. Другие функции логические и текстовые значения игнорируют.

4.5.1. Функции ДИСП и ДИСПА

Эти функции вычисляют выборочную дисперсию по выборке x_1, x_2, \dots, x_n (которая задается аргументами Число) по формуле

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \text{ где } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Это несмещенная оценка неизвестной дисперсии распределения генеральной совокупности.

4.5.2. Функции ДИСПР и ДИСПРА

Эти функции вычисляют выборочную дисперсию по выборке x_1, x_2, \dots, x_n (которая задается аргументами Число) по формуле

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \text{ где } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

Это асимптотически несмещенная оценка неизвестной дисперсии распределения генеральной совокупности.

4.5.3. Функция КВАДРОТКЛ

Эта функция по выборке x_1, x_2, \dots, x_n (которая задается аргументами Число) вычисляет сумму квадратов отклонений выборочных значений от выборочного среднего, т.е. вычисляет величину $\sum_{i=1}^n (x_i - \bar{x})^2$, где $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

Значение, возвращаемое этой функцией, можно использовать для вычисления выборочной дисперсии или выборочного среднеквадратического отклонения.

4.5.4. Функции СТАНДОТКЛОН и СТАНДОТКЛОНА

Эти функции вычисляют выборочное среднеквадратическое (стандартное) отклонение по выборке x_1, x_2, \dots, x_n (которая задается аргументами Число) по формуле

$$s_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}, \text{ где } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

4.5.5. Функции СТАНДОТКЛОНП и СТАНДОТКЛОНПА

Эти функции вычисляют выборочное среднееквадратическое (стандартное) отклонение по выборке x_1, x_2, \dots, x_n (которая задается аргументами Число) по формуле

$$S_n = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}, \text{ где } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i.$$

4.5.6. Функция СРОТКЛ

Эта функция по выборке x_1, x_2, \dots, x_n (которая задается аргументами Число) вычисляет среднее арифметическое модулей отклонений выборочных значений от выборочного среднего, т.е. величину $\frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|$, где $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

Значение, возвращаемое этой функцией, показывает (наряду со среднееквадратическим отклонением) меру рассеивания распределения генеральной совокупности относительно математического ожидания.

4.6. Функции для вычисления значений функций распределения

Это, по-видимому, самая обширная группа статистических функций. В нее входят такие функции.

<i>Функция</i>	<i>Назначение</i>
ГРАСП	Возвращает значения функции F-распределения
БЕТАРАСП	Возвращает значения функции бета-распределения
БИНОМРАСП	Возвращает значения биномиального распределения
ВЕЙБУЛЛ	Возвращает значения распределения Вейбулла-Гнеденко
ГАММАРАСП	Возвращает значения гамма-распределения
ГИПЕРГЕОМЕТ	Возвращает значения гипергеометрического распределения
ЛОГНОРМРАСП	Возвращает значения логарифмически нормального распределения
НОРМРАСП	Возвращает значения нормального распределения
НОРМСТРАСП	Возвращает значения стандартного нормального распределения
ОТРБИНОМРАСП	Возвращает значения отрицательного биномиального распределения
ПУАССОН	Возвращает значения распределения Пуассона
СТЬЮДРАСП	Возвращает значения распределения Стьюдента
ХИ2РАСП	Возвращает значения распределения χ^2
ЭКСПРАСП	Возвращает значения экспоненциального распределения

4.6.1. Функция FРАСП

Эта функция используется в статистическом анализе для проверки статистических гипотез. Она вычисляет вероятность $P(X \geq x)$, где X — случайная величина, имеющая F -распределение (распределение Снедекора) с (m, n) степенями свободы (см. раздел 1.5.7)¹. Чтобы с помощью этой функции вычислить значение функции F -распределения $F(u)$, необходимо применить формулу $=1 - \text{FРАСП}(u; m; n)$ (m и n — заданные значения степеней свободы), как показано на рис. 4.6.

Синтаксис функции:

FРАСП(х;Степень_свободы1;Степень_свободы2)

Здесь x — это значение, для которого вычисляется функция, Степень_свободы1 и Степень_свободы2 — значения степеней свободы F -распределения. Если какое-либо из этих значений не целое, то берется целая часть этого значения.

Если какой-либо из аргументов не является числом, то функция возвращает значение ошибки #ЗНАЧ!. Если x отрицательно либо если значения степеней свободы меньше 1 или больше 10^{10} , то функция возвращает значение ошибки #ЧИСЛО!.

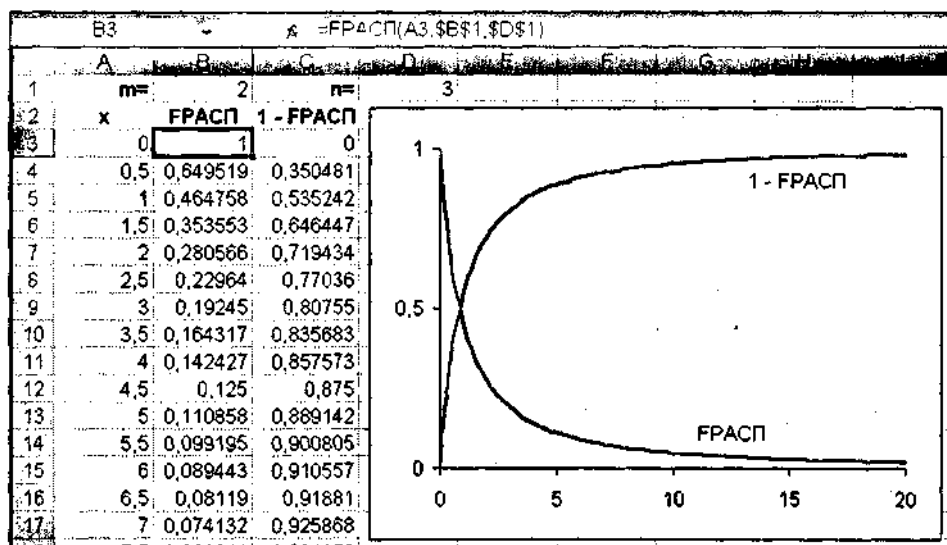


Рис. 4.6. Графики функций FРАСП и 1 - FРАСП

4.6.2. Функция БЕТАРАСП

Эта функция вычисляет значения функции бета-распределения с параметрами α и β , α и $\beta > 0$ (см. раздел 1.5.9).

Синтаксис функции:

БЕТАРАСП(х;Альфа;Бета;А;В)

¹ Отметим, что в справочной системе Excel при описании этой функции ошибочно указано, что она вычисляет вероятность $P(X < x)$ (X — случайная величина, имеющая F -распределение). Эта ошибка повторяется в большинстве книг, содержащих описание статистических функций.

Аргумент x — это значение из интервала от A до B , для которого вычисляется функция. Альфа и Бета — параметры распределения. Необязательные аргументы A и B — соответственно нижняя и верхняя границы интервала изменения x . Если аргументы A и B опущены, то по умолчанию принимается, что $A = 0$ и $B = 1$.

Если какой-либо из аргументов не является числом, то функция БЕТАРАСП возвращает значение ошибки #ЗНАЧ!. Если Альфа или Бета ≤ 0 либо если $x < A$, $x > B$ или $A = B$, то функция возвращает значение ошибки #ЧИСЛО!.

4.6.3. Функция БИНОМРАСП

Напомним (см. раздел 1.4.3), что биномиальное распределение часто рассматривают как модель случайных экспериментов, состоящих из n независимых одинаковых испытаний, в результате каждого из которых с вероятностью p может произойти исход “1” и с вероятностью $(1 - p)$ — исход “0”. Тогда случайная величина, равная количеству k исходов “1” в n испытаниях, имеет биномиальное распределение. Функция БИНОМРАСП позволяет вычислять как значения вероятностей $P(X = k)$ при любых n , p и k , так и значения функции распределения $F(x)$.

Синтаксис функции:

БИНОМРАСП(Число_успехов; Число_испытаний; Вероятность_успеха; Интегральная)

Здесь аргумент Число_успехов — это количество испытаний k , в которых произошел исход “1”. Число_испытаний — количество испытаний n . Вероятность_успеха — вероятность p исхода “1”. Аргумент Интегральная принимает логическое значение: если этот аргумент имеет значение ИСТИНА (или 1), то функция возвращает значение функции распределения, т.е. вероятность того, что число исходов “1” не менее значения аргумента Число_успехов; если этот аргумент имеет значение ЛОЖЬ (или 0), то вычисляется вероятность того, что число исходов “1” в точности равно значению аргумента Число_успехов.

Если значения аргументов Число_успехов и Число_испытаний не являются целыми числами, то в качестве аргументов берется целая часть этих чисел. Если первые три аргумента не являются числами, то функция возвращает значение ошибки #ЗНАЧ!. Если значение аргумента Число_успехов отрицательно или больше значения аргумента Число_испытаний, то функция возвращает значение ошибки #ЧИСЛО!. Функция возвращает такую же ошибку, если значение аргумента Вероятность_успеха не принадлежит интервалу $(0, 1)$.

4.6.4. Функция ВЕЙБУЛЛ

Данная функция может вычислять как значения плотности вероятности, так и значения функции распределения Вейбулла–Гнеденко (см. раздел 1.5.11).

Синтаксис функции:

ВЕЙБУЛЛ(x ; Альфа; Бета; Интегральная)

Аргумент x — значение, для которого вычисляется функция. Альфа и Бета — неотрицательные параметры распределения. Аргумент Интегральная принимает логическое значение: если этот аргумент имеет значение ИСТИНА (или 1), то функция возвращает значение функции распределения; если этот аргумент имеет значение ЛОЖЬ (или 0), то вычисляется значение функции плотности вероятности.

Если первые три аргумента не являются числами, то функция возвращает значение ошибки #ЗНАЧ!. Если значения этих аргументов отрицательны, то функция возвращает значение ошибки #ЧИСЛО!.

4.6.5. Функция ГАММАРАСП

Данная функция может вычислять как значения плотности вероятности, так и значения функции гамма-распределения (см. раздел 1.5.10).

Синтаксис функции:

ГАММАРАСП(х;Альфа;Бета;Интегральная)

Аргумент х — значение, для которого вычисляется функция. Альфа и Бета — неотрицательные параметры распределения. Аргумент Интегральная принимает логическое значение: если этот аргумент имеет значение ИСТИНА (или 1), то функция возвращает значение функции распределения; если этот аргумент имеет значение ЛОЖЬ (или 0), то вычисляется значение функции плотности вероятности.

Если первые три аргумента не являются числами, то функция возвращает значение ошибки #ЗНАЧ!. Если значения этих аргументов отрицательны, то функция возвращает значение ошибки #ЧИСЛО!.

4.6.6. Функция ГИПЕРГЕОМЕТ

Данная функция вычисляет вероятности $P(X = k)$, где случайная величина X имеет гипергеометрическое распределение с параметрами N , n и p ($N \geq n \geq 0$, $0 < p < 1$) (см. раздел 1.4.6).

Синтаксис функции:

ГИПЕРГЕОМЕТ(Число_успехов_в_выборке;Размер_выборки;Число_успехов_в_совокупности;Размер_совокупности)

Аргумент Число_успехов_в_выборке — это значение k , аргумент Размер_выборки — значение n , Число_успехов_в_совокупности — значение pN , Размер_совокупности — это значение N .

Функция ГИПЕРГЕОМЕТ выполняет вычисления по формуле

$$P(X = k) = \frac{C_{np}^k C_{N(1-p)}^{n-k}}{C_N^n}, \quad k = 0, 1, 2, \dots, n,$$

где C_n^k — биномиальный коэффициент.

Все аргументы функции округляются до ближайших целых, не превышающих заданных значений аргументов. Если какой-либо аргумент не является числом, то функция возвращает значение ошибки #ЗНАЧ!. Если значение аргумента Число_успехов_в_выборке отрицательно или превосходит меньшее из чисел Размер_выборки и Число_успехов_в_совокупности, функция возвращает значение ошибки #ЧИСЛО!. Если аргумент Размер_выборки отрицательно или превосходит значение аргумента Размер_совокупности, функция также возвращает значение ошибки #ЧИСЛО!. Такие же ограничения накладываются на аргумент Число_успехов_в_совокупности. Значение аргумента Размер_совокупности должно быть положительным числом, иначе функция возвращает значение ошибки #ЧИСЛО!.

4.6.7. Функция ЛОГНОРМРАСП

Эта функция вычисляет значения функции логарифмически нормального распределения с параметрами σ и a^2 (см. раздел 1.5.8).

Синтаксис функции:

ЛОГНОРМРАСП(x ;Среднее;Стандартное_отклонение)

Аргумент x — значение, для которого вычисляется функция. Аргумент Среднее — это параметр σ , а Стандартное_отклонение — параметр σ .

Если какой-либо из аргументов не является числом, то функция возвращает значение ошибки #ЗНАЧ!. Если x или Стандартное_отклонение отрицательно или равно 0, то функция возвращает значение ошибки #ЧИСЛО!.

4.6.8. Функция НОРМРАСП

Эта функция вычисляет значения плотности вероятности и функции распределения для нормального распределения с параметрами σ и a^2 (см. раздел 1.5.4).

Синтаксис функции:

НОРМРАСП(x ;Среднее;Стандартное_отклонение;Интегральная)

Аргумент x — значение, для которого вычисляется функция. Аргументы Среднее и Стандартное_отклонение — параметры распределения σ и a соответственно. Аргумент Интегральная принимает логическое значение: если этот аргумент имеет значение ИСТИНА (или 1), то функция возвращает значение функции распределения; если этот аргумент имеет значение ЛОЖЬ (или 0), то вычисляется значение функции плотности вероятности.

Если первые три аргумента не являются числами, то функция возвращает значение ошибки #ЗНАЧ!. Если значение аргумента Стандартное_отклонение отрицательно или равно 0, то функция возвращает значение ошибки #ЧИСЛО!.

Если Среднее = 0 и Стандартное_отклонение = 1, то функция возвращает те же значения, что и функция НОРМСТРАСП.

4.6.9. Функция НОРМСТРАСП

Эта функция вычисляет значения функции распределения для стандартного нормального распределения (параметры $\sigma = 0$ и $\sigma^2 = 1$) (см. раздел 1.5.4).

Синтаксис функции:

НОРМСТРАСП(x)

Аргумент x — значение, для которого вычисляется функция. Если аргумент x не является числом, то функция возвращает значение ошибки #ЗНАЧ!.

4.6.10. Функция ОТРБИНОМРАСП

Данная функция вычисляет вероятность $P(X = k)$, где случайная величина X имеет отрицательное биномиальное распределение (распределение Паскаля) с параметрами r и p и $0 < p < 1$ (см. раздел 1.4.7). Эта вероятность вычисляется по формуле

$$P(X = k) = C_{r+k-1}^k p^r (1-p)^k, \quad k = 0, 1, 2, \dots,$$

где C^* — биномиальный коэффициент.

Синтаксис функции:

ОТБИНОМРАСП(Число_k;Число_r;Вероятность)

Аргумент Число_k принимает значение k , аргумент Число_r — значение параметра r и Вероятность — значение вероятности p .

Значения первых двух аргументов функции округляются до ближайших целых, не превышающих заданных значений аргументов. Если какой-либо аргумент не является числом, то функция возвращает значение ошибки #ЗНАЧ!. Если значение аргумента Вероятность выходит за интервал $(0, 1)$, функция возвращает значение ошибки #ЧИСЛО!. Если сумма первых двух аргументов меньше 1, функция возвращает то же значение ошибки #ЧИСЛО!.

4.6.11. Функция ПУАССОН

Функция вычисляет значения распределения Пуассона с параметром λ , $\lambda > 0$ (см. раздел 1.4.4). Это распределение является моделью для описания случайного числа появления определенных событий в фиксированный промежуток времени или в фиксированной области пространства.

Синтаксис функции:

Пуассон(х;Среднее;Интегральная)

Аргумент x — количество событий. Аргумент Среднее — значение параметра λ . Аргумент Интегральная принимает логическое значение: если этот аргумент имеет значение ИСТИНА (или 1), то функция вычисляет значение функции распределения, т.е. вероятность того, что число случайных событий будет от 0 до x включительно; если этот аргумент имеет значение ЛОЖЬ (или 0), то вычисляется вероятность того, что событий будет в точности x .

Если x — не целое число, то в качестве аргумента берется целая часть этого числа. Если первые два аргумента функции не являются числами, то функция возвращает значение ошибки #ЗНАЧ!. Если x и Среднее отрицательны или равны 0, то функция возвращает значение ошибки #ЧИСЛО!.

4.6.12. Функция СТЬЮДРАСП

Эта функция используется в статистическом анализе для проверки статистических гипотез. В зависимости от значения аргумента Хвосты она вычисляет либо вероятность $P(X \geq x)$, где X — случайная величина, имеющая распределение Стьюдента с n степенями свободы (см. раздел 1.5.6), либо вероятность $1 - P(|X| \leq x) = P(X \leq -x) + P(X \geq x)$. (В силу симметрии распределения Стьюдента во втором случае значение, возвращаемое функцией, будет в два раза больше чем значение, возвращаемое в первом случае.) Чтобы с помощью этой функции вычислить значение функции распределения $F(u)$, необходимо применить формулу $=1 - \text{СТЬЮДРАСП}(u;n;1)$ (n — заданное значение степени свободы, 1 — значение аргумента Хвосты).

Синтаксис функции:

СТЬЮДРАСП(х;Степень_свободы;Хвосты)

Здесь x — неотрицательное значение, для которого вычисляется функция, Степень_свободы — значение степени свободы распределения. Аргумент Хвосты

может принимать значение 1 или 2: если этот аргумент равен 1, то функция возвращает значение вероятности $P(X \geq x)$; если же аргумент равен 2, то функция возвращает значение вероятности $P(X \leq -x) + P(X \geq x)$. Если какое-либо из значений последних двух аргументов не целое, то берется целая часть этого значения.

Если какой-либо из аргументов не является числом, то функция возвращает значение ошибки #ЗНАЧ!. Если x отрицательно либо если значение степени свободы меньше 1 или значение аргумента Хвосты выходит за интервал (1, 3), то функция возвращает значение ошибки #ЧИСЛО!.

4.6.13. Функция ХИ2РАСП

Эта функция используется в статистическом анализе для проверки статистических гипотез. Она вычисляет вероятность $P(X \geq x)$, где X — случайная величина, имеющая распределение χ^2 с n степенями свободы (см. раздел 1.5.5). Чтобы с помощью этой функции вычислить значение функции распределения $F(u)$, необходимо применить формулу $=1 - \text{ХИ2РАСП}(u;n)$ (n — заданное значение степеней свободы).

Синтаксис функции:

ХИ2РАСП(х;Степень_свободы)

Здесь x — значение, для которого вычисляется функция, Степень_свободы — значение степеней свободы распределения. Если значение аргумента Степень_свободы не целое, то берется целая часть этого значения.

Если какой-либо из аргументов не является числом, то функция возвращает значение ошибки #ЗНАЧ!. Если x отрицательно, либо если значение степеней свободы меньше 1 или больше 10^{10} , то функция возвращает значение ошибки #ЧИСЛО!.

4.6.14. Функция ЭКСПРАСП

Эта функция вычисляет значения плотности вероятности и функции распределения для экспоненциального (показательного) распределения с параметром λ , $\lambda > 0$ (см. раздел 1.5.3).

Синтаксис функции:

ЭКСПРАСП(х;Лямбда;Интегральная)

Аргумент x — значение, для которого вычисляется функция. Аргумент Лямбда — параметр распределения λ . Аргумент Интегральная принимает логическое значение: если этот аргумент имеет значение ИСТИНА (или 1), то функция возвращает значение функции распределения; если этот аргумент имеет значение ЛОЖЬ (или 0), то вычисляется значение функции плотности вероятности.

Если первые два аргумента не являются числами, то функция возвращает значение ошибки #ЗНАЧ!. Если значение аргумента x отрицательно либо если значение аргумента Лямбда меньше или равно 0, то функция возвращает значение ошибки #ЧИСЛО!.

4.7. Функции, обратные к функциям распределения

В эту группу входят функции, вычисляющие значения функций, обратных к функциям распределения.

Функция	Назначение
ФРАСПОБР	Возвращает обратное значение для F -распределения
БЕТАОБР	Вычисляет значение функции, обратной к функции распределения бета-распределения
ГАММАОБР	Вычисляет значение функции, обратной к функции распределения гамма-распределения
КРИТБИНОМ	Вычисляет значение функции, обратной к функции распределения биномиального распределения
ЛОГНОРМОВР	Вычисляет значение функции, обратной к функции распределения логарифмически нормального распределения
НОРМОВР	Вычисляет значение функции, обратной к функции нормального распределения
НОРМСТОБР	Вычисляет значение функции, обратной к функции стандартного нормального распределения
СТЮДРАСПОВР	Вычисляет значение функции, обратной к функции распределения Стюдента
ХИ2ОВР	Вычисляет значение функции, обратной к функции распределения χ^2

Эти функции имеют обязательный аргумент Вероятность (и, конечно, аргументы, задающие параметры распределения), в соответствии с которым вычисляется значение функции. Обращаем внимание на то, что не все из этих функций вычисляют значения функций, обратных к функциям распределения. Если определение значения обратной функции эквивалентно решению уравнения $P(X \leq u) = p$, где X — случайная величина, имеющая данное распределение, p — заданная вероятность, а u — искомая величина (т.е. $u = F^{-1}(p)$, F^{-1} — функция, обратная к функции распределения $F(u) = P(X \leq u)$), то некоторые функции из этой группы решают уравнение $P(X \geq u) = p$. Чтобы в этом случае найти значение обратной функции, необходимо вычислить статистическую функцию этого типа с аргументом Вероятность = $1 - p$. Такие функции удобно использовать для построения критических областей критериев проверки гипотез. Функции, вычисляющие значения обратных функций, удобно применять для моделирования случайных величин, имеющих заданное распределение.

4.7.1. Функция ФРАСПОБР

Это функция, вычисляющая корень уравнения $P(X \geq u) = p$, где X — случайная величина, имеющая F -распределение (распределение Снедекора) с (m, n) степенями свободы ($m, n \geq 1$) (см. раздел 1.5.7).

Синтаксис функции:

ФРАСПОБР(Вероятность;Степень_свободы1;Степень_свободы2)

Аргумент Вероятность — это значение вероятности p . Аргументы Степень_свободы1 и Степень_свободы2 — значения степеней свободы, т.е. параметры m и n . Если значение какого-либо из последних аргументов не является целым числом, берется целая часть этого числа.

Если какой-либо из аргументов не является числом, то функция возвращает значение ошибки #ЗНАЧ!. Если значение аргумента Вероятность не принадлежит интервалу $(0, 1)$ либо если значения аргументов Степень_свободы1 и Степень_свободы2 меньше 1 или больше 10^{10} , то функция возвращает значение ошибки #ЧИСЛО!.

4.7.2. Функция БЕТАОБР

Данная функция возвращает значение функции, обратной к функции бета-распределения с параметрами α и β ($\alpha > 0$, $\beta > 0$) (см. раздел 1.5.9).

Синтаксис функции:

БЕТАОБР(Вероятность;Альфа;Бета;А;В)

Аргумент Вероятность — это значение вероятности p . Аргументы Альфа и Бета — неотрицательные параметры распределения. Необязательные аргументы А и В задают соответственно нижнюю и верхнюю границы интервала изменения случайной величины. Если значения этих аргументов не заданы, то по умолчанию принимается, что $A = 0$ и $B = 1$.

Если какой-либо из аргументов не является числом, то функция возвращает значение ошибки #ЗНАЧ!. Если значения аргументов Альфа и Бета меньше или равны 0 либо если значение аргумента Вероятность выходит за интервал $(0, 1)$, то функция возвращает значение ошибки #ЧИСЛО!.

4.7.3. Функция ГАММАОБР

Эта функция возвращает значение функции, обратной к функции гамма-распределения с параметрами α и λ ($\alpha > 0$, $\lambda > 0$) (см. раздел 1.5.10).

Синтаксис функции:

ГАММАОБР(Вероятность;Альфа;Бета)

Аргумент Вероятность — это значение вероятности p . Аргументы Альфа и Бета — неотрицательные параметры распределения, при этом параметр Бета равен $1/\lambda$.

Если какой-либо из аргументов не является числом, то функция возвращает значение ошибки #ЗНАЧ!. Если значения аргументов Альфа и Бета меньше или равны 0 либо если значение аргумента Вероятность выходит за интервал $(0, 1)$, то функция возвращает значение ошибки #ЧИСЛО!.

4.7.4. Функция ЛОГНОРМОБР

Эта функция возвращает значение функции, обратной к функции логарифмически нормального распределения с параметрами m и σ^2 (см. раздел 1.5.8).

Синтаксис функции:

ЛОГНОРМОБР(Вероятность;Среднее;Стандартное_отклонение)

Аргумент Вероятность — это значение вероятности p . Аргументы Среднее и Стандартное_отклонение — параметры распределения m и σ .

Если какой-либо из аргументов не является числом, то функция возвращает значение ошибки #ЗНАЧ!. Если значение аргумента Стандартное_отклонение меньше или равно 0 либо если значение аргумента Вероятность выходит за интервал $(0, 1)$, то функция возвращает значение ошибки #ЧИСЛО!.

4.7.5. Функция НОРМОБР

Функция возвращает значение функции, обратной к функции нормального распределения с параметрами m и σ^2 (см. раздел 1.5.4).

Синтаксис функции:

НОРМОБР(Вероятность;Среднее;Стандартное_отклонение)

Аргумент Вероятность — это значение вероятности p . Аргументы Среднее и Стандартное_отклонение — параметры распределения m и σ .

Если какой-либо из аргументов не является числом, то функция возвращает значение ошибки #ЗНАЧ!. Если значение аргумента Стандартное_отклонение меньше или равно 0 либо если значение аргумента Вероятность выходит за интервал (0, 1), то функция возвращает значение ошибки #ЧИСЛО!.

4.7.6. Функция НОРМСТОБР

Функция возвращает значение функции, обратной к функции стандартного нормального распределения (в этом случае $m = 0$ и $\sigma^2 = 1$) (см. раздел 1.5.4).

Синтаксис функции:

НОРМ'СТОБР(Вероятность)

Аргумент Вероятность — значение вероятности p .

Если аргумент не является числом, то функция возвращает значение ошибки #ЗНАЧ!. Если значение аргумента Вероятность выходит за интервал (0, 1), функция возвращает значение ошибки #ЧИСЛО!.

4.7.7. Функция СТЬЮДРАСПОБР

Это функция, вычисляющая корень уравнения $P(X \geq u) = p$, где X — случайная величина, имеющая распределение Стьюдента с n степенями свободы ($n \geq 1$) (см. раздел 1.5.6).

Синтаксис функции:

СТЬЮДРАСПОБР(Вер'оятность;Степень_свободы)

Аргумент Вероятность — это значение вероятности p . Аргумент Степень_свободы — значение степени свободы, т.е. параметр n . Если значение этого аргумента не является целым числом, берется целая часть этого числа.

Если какой-либо из аргументов не является числом, то функция возвращает значение ошибки #ЗНАЧ!. Если значение аргумента Вероятность не принадлежит интервалу (0, 1) либо если значение аргумента Степень_свободы меньше 1 или больше 10^{10} , то функция возвращает значение ошибки #ЧИСЛО!.

4.7.8. Функция ХИ2ОБР

Функция вычисляет корень уравнения $P(X \geq u) = p$, где X — случайная величина, имеющая распределение χ^2 с n степенями свободы ($n \geq 1$) (см. раздел 1.5.5).

Синтаксис функции:

ХИ2ОБР(Вероятность;Степень_свободы)

Аргумент Вероятность — это значение вероятности p . Аргумент Степень_свободы — значение степени свободы, т.е. параметр n . Если значение этого аргумента не является целым числом, берется целая часть этого числа.

Если какой-либо из аргументов не является числом, то функция возвращает значение ошибки #ЗНАЧ!. Если значение аргумента Вероятность не принадлежит интервалу (0, 1) либо если значение аргумента Степень_свободы меньше 1 или больше 10^{10} , то функция возвращает значение ошибки #ЧИСЛО!.

4.7.9. Функция КРИТБИНОМ

Функция возвращает значение функции, обратной к функции биномиального распределения с параметрами n и p ($0 < p < 1$, $n \geq 1$). Напомним (см. раздел 1.4.3), что биномиальное распределение является моделью случайных экспериментов, состоящих из n независимых одинаковых испытаний, и в результате каждого из них с вероятностью p может произойти исход "1" и с вероятностью $(1 - p)$ — исход "0". Тогда случайная величина X , равная количеству k исходов "1" в n испытаниях, имеет биномиальное распределение. Функция КРИТБИНОМ вычисляет наименьшее значение k , при котором $P(X = k) \geq \alpha$ (α — заданное число).

Синтаксис функции:

КРИТБИНОМ(Число_испытаний;Вероятность;Альфа)

Аргумент Число_испытаний — количество независимых испытаний n . Если значение этого аргумента — не целое число, то берется целая часть этого числа. Аргумент Вероятность — вероятность p исхода "1" в каждом испытании, т.е. параметр распределения. Аргумент Альфа — значение вероятности α .

Если какой-либо из аргументов не является числом, то функция возвращает значение ошибки #ЗНАЧ!. Если значение Число_испытаний отрицательно либо если значения аргументов Вероятность и Альфа выходят за интервал (0, 1), то функция возвращает значение ошибки #ЧИСЛО!.

4.8. Функции для проверки статистических критериев

Функции этой группы выполняют расчеты для различных статистических критериев.

Функция	Назначение
ZТЕСТ	Используется для проверки гипотез о значении математического ожидания нормально распределенной генеральной совокупности
TТЕСТ	Используется для проверки гипотез о равенстве (неравенстве) математических ожиданий двух выборок (критерий Стьюдента)
FТЕСТ	Используется для проверки гипотез о равенстве (неравенстве) дисперсий двух выборок
XI2ТЕСТ	Используется для проверки гипотез о принадлежности выборки определенному классу распределений (критерий χ^2)

4.8.1. Функция ZTEST

Эта функция используется для проверки гипотез о значении неизвестного математического ожидания генеральной совокупности, распределенной по нормальному закону, при известной дисперсии распределения. Чтобы пояснить вычисления, выполняемые функцией ZTEST, напомним статистическую модель и проверяемые гипотезы (см. раздел 2.4.1).

Статистическая модель. Выборка x_1, x_2, \dots, x_n получена из генеральной совокупности с нормальным законом распределения и с неизвестным математическим ожиданием μ и известной дисперсией σ^2 .

Гипотезы

а) Равенство

$$H_0: \mu = m_0$$

$$H_1: \mu \neq m_0$$

б) Неравенство

$$H_0: \mu \leq m_0$$

$$H_1: \mu > m_0$$

в) Неравенство

$$H_0: \mu \geq m_0$$

$$H_1: \mu < m_0$$

Здесь m_0 — заданное число. Задан уровень значимости α .

Функция ZTEST сначала вычисляет значение критериальной статистики

$$T = \frac{\sqrt{n}(\bar{x} - m_0)}{\sigma}, \text{ где } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \text{ а затем — вероятность } ZTEST = 1 - F(T), \text{ где } F —$$

функция распределения стандартного нормального закона.

Случай а). Гипотеза H_0 принимается, если выполняется неравенство $\alpha/2 \leq ZTEST \leq 1 - \alpha/2$, иначе гипотеза H_0 отклоняется.

Случай б). Гипотеза H_0 принимается, если $ZTEST \leq 1 - \alpha$.

Случай в). Гипотеза H_0 принимается, если $\alpha \leq ZTEST$.

Синтаксис функции:

$$ZTEST(\text{Массив}; x; \text{Сигма})$$

Аргумент Массив — массив данных или адрес диапазона ячеек, содержащий выборочные значения x_1, x_2, \dots, x_n . Аргумент x — проверяемое значение математического ожидания (т.е. значение m_0). Необязательный аргумент Сигма — значение стандартного отклонения σ генеральной совокупности. Если этот аргумент опущен, то используется выборочное стандартное отклонение. (Но поскольку в этом случае все равно используется функция распределения нормального закона, в таком варианте функцию ZTEST можно использовать только при достаточно большом объеме выборки.)

Если аргумент Массив пуст, то функция возвращает значение ошибки #Н/Д.

4.8.2. Функция TTEST

Эта функция используется для проверки гипотезы о равенстве (неравенстве) неизвестных математических ожиданий двух генеральных совокупностей, распределенных по нормальному закону, причем функция работает как для зависимых выборок, так и для независимых и при условиях равенства и неравенства дисперсий выборок.

Чтобы пояснить вычисления, выполняемые функцией TTEST, приведем соответствующие статистические модели и проверяемые гипотезы (см. раздел 2.4.2).

Статистическая модель 1. Двумерная выборка $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ объемом n извлечена из двумерной нормальной совокупности с неизвестными

математическими ожиданиями соответственно μ_1 и μ_2 компонентов выборки. Этой модели в функции ТТЕСТ соответствует значение 1 аргумента Тип.

Статистическая модель 2. Выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_m объемом соответственно n и m извлечены из совокупностей, имеющих нормальные распределения с равными дисперсиями σ_1^2 и σ_2^2 и математическими ожиданиями μ_1 и μ_2 соответственно. Этой модели в функции ТТЕСТ соответствует значение 2 аргумента Тип.

Статистическая модель 3. Выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_m объемом соответственно n и m извлечены из совокупностей, имеющих нормальные распределения с неравными дисперсиями σ_1^2 и σ_2^2 и математическими ожиданиями μ_1 и μ_2 соответственно. Этой модели в функции ТТЕСТ соответствует значение 3 аргумента Тип.

Во всех статистических моделях проверяются следующие гипотезы.

Гипотезы

а) Равенство

б) Неравенство

$H_0: \mu_1 = \mu_2$

$H_0: \mu_1 \leq \mu_2$

$H_1: \mu_1 \neq \mu_2$

$H_1: \mu_1 > \mu_2$

Задан уровень значимости α .

Синтаксис функции:

ТТЕСТ(Массив1, Массив2, Хвосты, Тип)

Аргумент Массив1 представляет первую выборку x_1, x_2, \dots, x_n , аргумент Массив2 — вторую выборку y_1, y_2, \dots, y_m . Значение аргумента Хвосты равно 1 для проверки гипотезы о неравенстве математических ожиданий и равно 2 для проверки гипотезы о равенстве. Аргумент Тип должен иметь значение 1 для статистической модели 1, значение 2 для модели 2 и значение 3 для модели 3.

В зависимости от статистической модели функция ТТЕСТ выполняет такие вычисления.

Статистическая модель 1 (значение аргумента Тип равно 1). Вычисляются n разностей $d_1 = x_1 - y_1, d_2 = x_2 - y_2, \dots, d_n = x_n - y_n$ и по ним определяются среднее $\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$ и выборочная дисперсия разностей $S_n^2 = \frac{1}{n} \sum_{i=1}^n (d_i - \bar{d})^2$. По вычис-

ленному значению критериальной статистики $T = \frac{\bar{d}}{S_n / \sqrt{n}}$ функция ТТЕСТ воз-

вращает значения $1 - F(T)$, если значение аргумента Хвосты равно 1, или $1 - F(T) + F(-T)$, если значение аргумента Хвосты равно 2, где $F(x)$ — функция распределения Стьюдента с $(n - 1)$ степенью свободы.

Гипотезы о равенстве и неравенстве принимаются, если значение, возвращаемое функцией ТТЕСТ, больше заданного уровня значимости α . Напомним, что для проверки гипотезы о равенстве значение аргумента Хвосты равно 2, а для проверки гипотезы о неравенстве значение аргумента Хвосты равно 1.

В Excel этот критерий реализует средство Парный двухвыборочный t-тест для средних из пакета анализа (см. раздел 5.9).

Статистическая модель 2 (значение аргумента Тип равно 2). По каждой выборке вычисляются выборочные средние и выборочные дисперсии: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$,

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad \bar{y} = \frac{1}{m} \sum_{i=1}^m y_i, \quad S_y^2 = \frac{1}{m} \sum_{i=1}^m (y_i - \bar{y})^2. \text{ По вычисленному значению}$$

критериальной статистики $T = \frac{\sqrt{n+m-2}(\bar{x} - \bar{y})}{\sqrt{\frac{n+m}{nm} \sqrt{(n-1)S_x^2 + (m-1)S_y^2}}}$ функция ТТЕСТ воз-

вращает значения $1 - F(T)$, если значение аргумента Хвосты равно 1, или $1 - F(T) + F(-T)$, если значение аргумента Хвосты равно 2, где $F(x)$ — функция распределения Стьюдента с $(n + m - 2)$ степенью свободы.

Гипотезы о равенстве и неравенстве принимаются, если значение, возвращаемое функцией ТТЕСТ, больше заданного уровня значимости α .

В Excel этот критерий реализует средство Двухвыборочный t-тест с одинаковыми дисперсиями из пакета анализа (см. раздел 5.7).

Статистическая модель 3 (значение аргумента Тип равно 3). По каждой выборке вычисляются выборочные средние и выборочные дисперсии: $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$,

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad \bar{y} = \frac{1}{m} \sum_{i=1}^m y_i, \quad S_y^2 = \frac{1}{m} \sum_{i=1}^m (y_i - \bar{y})^2. \text{ По вычисленному значению}$$

критериальной статистики $T = \frac{\bar{x} - \bar{y}}{\sqrt{S_x^2/n + S_y^2/m}}$ функция ТТЕСТ возвращает значе-

ния $1 - F(T)$, если значение аргумента Хвосты равно 1, или $1 - F(T) + F(-T)$, если значение аргумента Хвосты равно 2, где $F(x)$ — функция распределения Стьюдента

со степенью свободы k , которая рассчитывается по формуле $k = \frac{(S_x^2/n + S_y^2/m)^2}{\frac{(S_x^2/n)^2}{n-1} + \frac{(S_y^2/m)^2}{m-1}}$.

Гипотезы о равенстве и неравенстве принимаются, если значение, возвращаемое функцией ТТЕСТ, больше заданного уровня значимости α .

В Excel этот критерий реализует средство Двухвыборочный t-тест с различными дисперсиями из пакета анализа (см. раздел 5.8).

На аргументы функции ТТЕСТ накладываются следующие ограничения. Если Тип = 1 (парный критерий), то Массив1 и Массив2 должны представлять выборки одинаковых объемов, иначе функция возвращает значение ошибки #Н/Д. В случае дробных значений аргументов Хвосты и Тип берется целая часть этих значений. Если значения этих аргументов не являются числами, функция возвращает значение ошибки #ЗНАЧ!. Если аргумент Хвосты имеет значение, отличное от 1 и 2, функция возвращает значение ошибки #ЧИСЛО!.

4.8.3. Функция ФТЕСТ

Эта функция реализует критерий Фишера проверки равенства дисперсий двух независимых выборок из нормально распределенных генеральных совокупностей (см. раздел 2.4.2). Напомним, что этот критерий реализуется при выполнении следующей статистической модели.

Статистическая модель. Выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_m объемом соответственно n и m извлечены из совокупностей, имеющих нормальные распределения с неизвестными дисперсиями σ_1^2 и σ_2^2 и математическими ожиданиями μ_1 и μ_2 соответственно.

При заданном уровне значимости α проверяется нулевая гипотеза $H_0: \sigma_1^2 = \sigma_2^2$ против альтернативной гипотезы $H_1: \sigma_1^2 \neq \sigma_2^2$.

Синтаксис функции:

ФТЕСТ(Массив1, Массив2)

Аргумент Массив1 представляет первую выборку x_1, x_2, \dots, x_n , аргумент Массив2 — вторую выборку y_1, y_2, \dots, y_m .

Функция выполняет следующие вычисления. Для каждой выборки вычисляются сначала выборочные дисперсии $S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$, $S_y^2 = \frac{1}{m} \sum_{i=1}^m (y_i - \bar{y})^2$, а за-

тем — их отношение $F = \frac{S_x^2}{S_y^2}$, которое является критериальной статистикой.

Функция ФТЕСТ возвращает значение $1 - F(F)$, где $F(x)$ — функция F -распределения со степенями свободы $k_1 = n - 1$ и $k_2 = m - 1$ (о F -распределении речь идет в разделе 1.5.7).

Нулевая гипотеза принимается, если значение, возвращаемое функцией, больше заданного уровня значимости α .

Если объем выборки, задаваемой аргументом Массив1 или Массив2, меньше двух либо если дисперсия одной из выборок равна нулю, то функция возвращает значение ошибки #ДЕЛ/0!

Этот критерий также реализует средство Двухвыборочный F -тест для дисперсий из пакета анализа (см. раздел 5.10).

4.8.4. Функция ХИ2ТЕСТ

Эта функция является частью критерия χ^2 проверки гипотез о принадлежности распределения выборки определенному классу распределений. Приведем статистическую модель для этого критерия (см. раздел 2.4.3).

Статистическая модель. Выборка, состоящая из независимых выборочных значений x_1, x_2, \dots, x_n , получена из генеральной совокупности, имеющей функцию распределения $F(u)$ и зависящей от m параметров, из которых m_1 параметров неизвестно.

Проверяется нулевая гипотеза H_0 : выборочные значения получены из генеральной совокупности с функцией распределения $F(u)$ (зависящей от m параметров, из которых m_1 параметров определяются по выборочным значениям) против альтернативной гипотезы H_1 : нулевая гипотеза неверна.

Задается уровень значимости α .

Чтобы проверить эти гипотезы, еще до применения функции ХИ2ТЕСТ необходимо провести следующие вычисления.

1. Область возможных выборочных значений разбить на k непересекающихся интервалов $\Delta_1 = (x^{(1)}, x^{(2)})$, $\Delta_2 = (x^{(2)}, x^{(3)})$, ..., $\Delta_k = (x^{(k)}, x^{(k+1)})$.
2. Подсчитать, сколько выборочных значений попало в каждый интервал Δ_i . Получаем ряд чисел n_1, n_2, \dots, n_k .

3. В предположении, что справедлива гипотеза H_0 , по формуле $v_i = n[F(x^{(i+1)}) - F(x^{(i)})]$ вычислить ожидаемое значение попаданий выборочных значений в каждый из интервалов Δ_i , где $x^{(i)}$ и $x^{(i+1)}$ — границы интервала Δ_i .

Итак, имеются два массива данных: $\{n_1, n_2, \dots, n_k\}$ и $\{v_1, v_2, \dots, v_k\}$.

Далее вступает в работу функция ХИ2ТЕСТ. По заданным массивам $\{n_1, n_2, \dots, n_k\}$ и $\{v_1, v_2, \dots, v_k\}$ она вычисляет критериальную статистику $T = \sum_{i=1}^k \frac{(n_i - v_i)^2}{v_i}$

и затем возвращает вероятность $P(X > T)$, где X — случайная величина, имеющая распределение χ^2 с $(k - 1)$ степенью свободы.

Если значение, возвращаемое функцией ХИ2ТЕСТ больше заданного уровня значимости α , то гипотеза H_0 принимается. В противном случае гипотеза H_0 отклоняется.

Обращаем внимание, что функция ХИ2ТЕСТ использует распределение χ^2 с $(k - 1)$ степенью свободы, а не с $(k - m_1 - 1)$ степенью свободы. Поэтому критерий, выполняемый с помощью этой функции, имеет большую вероятность ошибки второго рода, т.е. большую вероятность принять нулевую гипотезу, если она неверна.

Синтаксис функции:

ХИ2ТЕСТ(Фактический_интервал;Ожидаемый_интервал)

Аргумент Фактический_интервал — это массив или ссылка на диапазон ячеек, содержащих числа n_1, n_2, \dots, n_k . Аргумент Ожидаемый_интервал — массив или ссылка на диапазон ячеек, содержащих числа v_1, v_2, \dots, v_k . Если аргументы содержат различные количества чисел, то функция возвращает значение ошибки #Н/Д.

Практическая реализация критерия χ^2 показана в главе 9, в разделе 9.3.

4.9. Функции для построения уравнения регрессии и прогнозирования

Функции этой группы весьма полезны при проведении регрессионного анализа.

Функция	Назначение
ЛГРФПРИБЛ	Возвращает параметры кривой, полученной в результате экспоненциальной аппроксимации
ЛИНЕЙН	Возвращает массив коэффициентов функции регрессии, полученный в результате аппроксимации исходных данных методом наименьших квадратов
НАКЛОН	Возвращает наклон прямой линейной регрессии
ОТРЕЗОК	Возвращает отрезок, отсекаемый на оси прямой линейной регрессии
ПРЕДСКАЗ	Возвращает предсказанное значение функции в точке X на основе линейной регрессии для массивов известных значений X и Y или интервалов данных
РОСТ	Рассчитывает прогнозируемый экспоненциальный рост на основании имеющихся данных
СТОШУХ	Возвращает для каждого значения X стандартную ошибку предсказанных значений Y (т.е. вычисленных значений функции регрессии)
ТЕНДЕНЦИЯ	Возвращает значение в соответствии с линейной функцией регрессии

Каждая из этих функций имеет не менее двух аргументов, один из которых задает массив значений независимой переменной X , а второй — массив значений зависимой переменной Y . В некоторых функциях можно задавать не только одномерный массив переменной X , но и двумерный, т.е. имеется возможность исследовать зависимость между векторной переменной X и скалярной Y и строить множественную регрессию. Функции ЛГРФПРИБЛ и РОСТ работают с экспоненциальной регрессией, остальные — с линейной. При построении уравнений регрессии все функции используют метод наименьших квадратов (см. раздел 3.4). Отметим, что другие средства Excel, в частности надстройка Пакет анализа (см. главу 5) и средства построения диаграмм (см. главу 6), имеют значительно большие возможности для построения и визуализации регрессионных зависимостей.

4.9.1. Функция ЛИНЕЙН

Применяя метод наименьших квадратов, данная функция рассчитывает коэффициенты линейной (относительно этих коэффициентов) регрессии, которая наилучшим образом аппроксимирует имеющиеся данные. Итак, имеется массив со значениями переменной X : одномерный $\{x_1, x_2, \dots, x_n\}$ (n — количество наблюдений), если исследуется зависимость переменной Y только от одной переменной, либо двумерный $\{x_{11}, x_{12}, \dots, x_{1n}, x_{21}, x_{22}, \dots, x_{2n}, \dots, x_{k1}, x_{k2}, \dots, x_{kn}\}$, если исследуется зависимость переменной Y от k переменных (т.е. переменная X в этом случае является вектором, состоящим из k компонентов: $X = (X_1, X_2, \dots, X_k)$). Задан также массив $\{y_1, y_2, \dots, y_n\}$ со значениями переменной Y . По этим данным мето-

дом наименьших квадратов строится уравнение линейной регрессии $\hat{Y} = mX + b$ в случае одномерной переменной X или $\hat{Y} = m_1X_1 + m_2X_2 + \dots + m_kX_k + b$ в случае, когда $X = (X_1, X_2, \dots, X_k)$. Функция ЛИНЕЙН по массивам исходных данных вычисляет коэффициенты m_i и b , а также может вычислить некоторые статистические характеристики этих коэффициентов и всего уравнения регрессии в целом.

Данную функцию можно также использовать для построения уравнения полиномиальной регрессии. Если массив значений X имеет структуру $\{x_1, x_2, \dots, x_n, x_1^2, x_2^2, \dots, x_n^2, \dots, x_1^k, x_2^k, \dots, x_n^k\}$, то в этом случае строится полиномиальная регрессия $\hat{Y} = m_1X + m_2X^2 + \dots + m_kX^k + b$.

Отметим, что функция возвращает массив значений коэффициентов m_i и b (не менее двух значений), поэтому функция должна задаваться в виде формулы массива (с использованием для ввода комбинации клавиш <Ctrl+Shift+Enter>), в противном случае (при вводе функции в одну ячейку) будет выведено значение только коэффициента m_1 .

Синтаксис функции:

ЛИНЕЙН(Значения_Y;Значения_X;Константа;Статистика)

Аргумент Значения_Y — одномерный массив (или ссылка на диапазон ячеек, содержащий этот массив) значений Y . Необязательный аргумент Значения_X — массив (или ссылка на диапазон ячеек, содержащий этот массив) значений X . Если данный аргумент опущен, предполагается, что это массив натуральных чисел {1; 2; 3; ...} такого же размера, как и массив Значения_Y. Если массив Значения_Y расположен в один столбец, то каждый столбец массива Значения_X интерпретируется как значения отдельной переменной X_i . Аналогично, если массив

Значения_Y расположен в одну строку, то каждая строка массива Значения_X интерпретируется как значения отдельной переменной X_i .

Аргумент Константа — логическое значение, которое указывает, должен ли коэффициент b быть равным 0. Если этот аргумент имеет значение ИСТИНА, 1 или опущен, то коэффициент b вычисляется как обычно. Если аргумент имеет значение ЛОЖЬ или 0, то b полагается равным 0 и значения коэффициентов m_i подбираются так, чтобы уравнение регрессии имело вид $\hat{Y} = m_1X_1 + m_2X_2 + \dots + m_kX_k$.

Аргумент Статистика принимает логическое значение, которое указывает, требуется ли рассчитывать дополнительные статистические характеристики регрессии. Если этот аргумент имеет значение ИСТИНА или 1, то функция рассчитывает и выводит эти дополнительные характеристики (см. таблицу, приведенную ниже; описание и пояснения к этим характеристикам даны в разделе 3.4.3). Если аргумент Статистика имеет значение ЛОЖЬ, 0 или опущен, то функция возвращает только значения коэффициентов m_i и b .

Таблица. Статистические характеристики, рассчитываемые функцией ЛИНЕЙН

Характеристика	Описание
s_1, s_2, \dots, s_k	Среднеквадратические отклонения для коэффициентов m_1, m_2, \dots, m_k
s_b	Среднеквадратическое отклонение для коэффициента b ($s_b = \#Н/Д$, если аргумент Константа имеет значение ЛОЖЬ)
R^2	Коэффициент детерминации. Сравниваются фактические значения Y и значения \hat{Y} , получаемые из уравнения регрессии; по результатам сравнения вычисляется коэффициент детерминации, нормированный от 0 до 1. Если он равен 1, то нет различия между фактическим и расчетными значениями Y . В противоположном случае, если коэффициент детерминации равен 0, уравнение регрессии плохо описывает значения Y
s_e	Остаточное среднеквадратическое отклонение
F	Критериальная статистика для проверки значимости уравнения регрессии
df	Степень свободы
SS_1	Сумма квадратов регрессии
SS_2	Сумма квадратов остатков

Отметим, что функция возвращает массив значений коэффициентов m_i и b (не менее двух значений), а также дополнительные статистические характеристики (если аргумент Статистика равен ИСТИНА). Поэтому функция должна задаваться в виде формулы массива, в противном случае (при вводе функции в одну ячейку) будет выведено значение только коэффициента m_k . В выходном массиве данные располагаются следующим образом.

m_k	m_{k-1}	...	m_2	m_1	b
s_k	s_{k-1}	...	s_2	s_1	s_b
R^2	s_e				
F	df				
SS_1	SS_2				

Остальные ячейки этого массива заполняются значениями #Н/Д.

Рассмотрим пример применения функции ЛИНЕЙН. Пусть, как показано на рис. 4.7, массив значений переменной X расположен в столбцах А и В, а массив значений переменной Y — в столбце С. Таким образом, переменная X — двумерная, имеет компоненты X_1 и X_2 . Выделим диапазон Е2:G6, в котором будут содержаться результаты вычислений. Вводим формулу =ЛИНЕЙН(С2:С17;А2:В17;;1) (см. рис. 4.7). Затем нажимаем комбинацию клавиш <Ctrl+Shift+Enter> (ввод формулы массива). Результат показан на рис. 4.8 (для удобства интерпретации результатов добавлены подписи к ячейкам).

ЛИНЕЙН						
=ЛИНЕЙН(С2:С17;А2:В17;;1)						
	А	В	С	Д	Е	Г
1	X1	X2	Y			
2	1,865	7,596	0,481		B17::1	
3	4,293	8,305	0,321			
4	2,580	9,023	5,496			
5	2,057	5,808	1,695			
6	2,720	3,180	8,784			
7	3,520	0,913	0,989			
8	8,667	6,586	5,924			
9	8,819	6,688	2,262			
10	4,761	2,996	7,077			
11	5,617	3,917	7,413			
12	3,316	2,905	5,271			
13	6,795	4,269	6,164			
14	5,324	5,911	4,646			
15	1,351	4,722	6,024			
16	9,775	1,507	8,821			
17	0,418	6,790	7,173			

Аргументы функции

ЛИНЕЙН

Известные_значения_y: C2:С17

Известные_значения_x: A2:В17

Конст: 1

Статистика: 1

Возвращает параметры линейного приближения по методу наименьших квадратов

Рис. 4.7. Ввод формулы

4.9.2. Функции НАКЛОН и ОТРЕЗОК

Эти функции вычисляют коэффициенты уравнения линейной регрессии $\hat{Y} = mX + b$, подсчитанные по методу наименьших квадратов (см. раздел 3.4.2): функция НАКЛОН вычисляет коэффициент m , функция ОТРЕЗОК — коэффициент b . (Чтобы сразу вычислить оба коэффициента, следует воспользоваться функцией ЛИНЕЙН.)

E2		=ЛИНЕЙН(C2:C17;A2:B17;1))					
	A	B	C				
1	X1	X2	Y		m2	m1	b
2	1,865	7,596	0,481		-0,43273	0,112059	6,599221
3	4,293	8,305	0,321	s2	0,3121	0,264707	2,280673
4	2,580	9,023	5,496	R2	0,159635	2,834828	#Н/Д
5	2,057	5,808	1,695	F	2,25539	13	#Н/Д
6	2,720	3,180	8,784	SS1			
7	3,520	0,913	0,989				
8	8,667	6,586	5,924				
9	8,819	6,688	2,262				
10	4,761	2,996	7,077				
11	5,617	3,917	7,413				
12	3,316	2,905	5,271				
13	6,795	4,269	6,164				
14	5,324	5,911	4,646				
15	1,351	4,722	6,024				
16	9,775	1,507	8,821				
17	0,418	6,790	7,173				

Рис. 4.8. Результаты вычислений

Синтаксис функций:

ФУНКЦИЯ(Значения_Y;Значения_X)

Аргумент Значения_Y — одномерный массив значений Y (или ссылка на диапазон ячеек, содержащий этот массив). Аргумент Значения_X — массив значений X (или ссылка на диапазон ячеек, содержащий этот массив).

Если аргументы содержат текст, логические значения или пустые ячейки, эти значения игнорируются; ячейки, содержащие нулевые значения, учитываются. Если аргументы пусты или содержат различные количества данных, то функции возвращают значение ошибки #Н/Д.

4.9.3. Функция СТОШУХ

Функция со странным названием СТОШУХ (кстати, последние буквы в этом названии — это не русское молодецкое “УХ-Х!”, а спокойные латинские буквы “игрек” и “икс”) вычисляет стандартную ошибку регрессии или корень квадратный из средней суммы остатков (см. раздел 3.4.3). Эту же величину вычисляет функция ЛИНЕЙН среди своих дополнительных статистических характеристик под названием s_e — остаточное среднеквадратическое отклонение.

Пусть имеется массив $\{x_1, x_2, \dots, x_n\}$ значений X и массив $\{y_1, y_2, \dots, y_n\}$ значений Y, по которым по методу наименьших квадратов строится уравнение линейной регрессии $\hat{Y} = mX + b$. Остатками называются разности $d_i = y_i - \hat{y}_i = y_i - mx_i - b$. Стандартная ошибка регрессии вычисляется по формуле

$s_y = \sqrt{\frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{y}_i)^2}$. Эта величина характеризует точность аппроксимации

исходных данных линейной функцией.

Синтаксис функции:

СТОШУХ(Значения_Y;Значения_X)

Аргумент Значения_Y — одномерный массив значений Y (или ссылка на диапазон ячеек, содержащий этот массив). Аргумент Значения_X — массив значений X (или ссылка на диапазон ячеек, содержащий этот массив).

Если аргументы содержат текст, логические значения или пустые ячейки, эти значения игнорируются; ячейки, содержащие нулевые значения, учитываются. Если аргументы пусты или содержат различные количества данных, то функции возвращают значение ошибки #Н/Д.

4.9.4. Функция ПРЕДСКАЗ

Пусть имеется массив $\{x_1, x_2, \dots, x_n\}$ значений X и массив $\{y_1, y_2, \dots, y_n\}$ значений Y, по которым методом наименьших квадратов строится уравнение линейной регрессии $\hat{Y} = mX + b$. Данная функция вычисляет значение $\hat{y} = mx + b$ для заданного значения x, т.е. “предсказывает” значение переменной Y, откуда и название функции.

Синтаксис функции:

ПРЕДСКАЗ(х;Значения_Y;Значения_X)

Аргумент x — значение, для которого вычисляется уравнение регрессии. Аргумент Значения_Y — одномерный массив значений Y (или ссылка на диапазон ячеек, содержащий этот массив). Аргумент Значения_X — массив значений X (или ссылка на диапазон ячеек, содержащий этот массив).

Если аргумент x не является числом, то функция возвращает значение ошибки #ЗНАЧ!. Если аргументы содержат текст, логические значения или пустые ячейки, эти значения игнорируются; ячейки, содержащие нулевые значения, учитываются. Если аргументы пусты или содержат различные количества данных, функции возвращают значение ошибки #Н/Д.

Эту функцию можно использовать для одновременного вычисления массива значений $\{\hat{y}\}$ по заданному массиву значений $\{x\}$, для чего в качестве аргумента x надо указать массив $\{x\}$, а саму функцию применить как формулу массива (нажав комбинацию клавиш <Ctrl+Shift+Enter>) к выделенному диапазону ячеек, в котором будет записан выходной массив значений $\{\hat{y}\}$.

4.9.5. Функция ТЕНДЕНЦИЯ

Эта функция, подобно предыдущей функции, вычисляет в соответствии с построенным методом наименьших квадратов уравнением регрессии значение \hat{Y} для конкретного значения X. Но в отличие от функции ПРЕДСКАЗ функция ТЕНДЕНЦИЯ может работать как с множественной линейной регрессией, так и с полиномиальной регрессией, что зависит от структуры содержимого входного массива значений переменной X.

Пусть задан массив $\{y_1, y_2, \dots, y_n\}$ со значениями переменной Y. Если массив значений переменной X является двумерным массивом вида $\{x_{11}, x_{12}, \dots, x_{1n}, x_{21}, x_{22}, \dots, x_{2n}, \dots, x_{k1}, x_{k2}, \dots, x_{kn}\}$, то в этом случае исследуется зависимость переменной Y от k переменных X_1, X_2, \dots, X_k и строится множественная линейная регрессия $\hat{Y} = m_1X_1 + m_2X_2 + \dots + m_kX_k + b$. Если же данный массив имеет структуру $\{x_1, x_2, \dots, x_n, x_1^2, x_2^2, \dots, x_n^2, \dots, x_1^k, x_2^k, \dots, x_n^k\}$, то в этом случае

строится полиномиальная регрессия $\hat{Y} = m_1X + m_2X^2 + \dots + m_kX^k + b$. Функция **ТЕНДЕНЦИЯ** по заданным значениям (x_1, x_2, \dots, x_k) и по уравнению множественной линейной регрессии или по значениям (x, x^2, \dots, x^k) и по уравнению полиномиальной регрессии вычисляет значение \hat{y} .

Если исходный массив значений X совпадает по размеру с массивом значений Y , то функция **ТЕНДЕНЦИЯ** для вычисления нового значения \hat{y} использует обычную линейную регрессию и в этом случае она не отличается от функции **ПРЕДСКАЗ**.

Синтаксис функции:

ТЕНДЕНЦИЯ(Значения_Y;Значения_X;Новые_значения_x;Константа)

Аргумент **Значения_Y** — одномерный массив значений Y (или ссылка на диапазон ячеек, содержащий этот массив). Аргумент **Значения_X** — массив значений X (или ссылка на диапазон ячеек, содержащий этот массив). Аргумент **Новые_значения_x** — значения, для которых вычисляется уравнение регрессии. Если аргумент **Значения_X** опущен, то предполагается, что это массив натуральных чисел {1; 2; 3; ...} такого же размера, как и массив аргумента **Значения_Y**. Если опущен аргумент **Новые_значения_x**, то по умолчанию предполагается, что он совпадает с аргументом **Значения_X**.

Аргумент **Константа** принимает логическое значение: если он имеет значение **ИСТИНА** или 1 либо опущен, то коэффициент уравнения регрессии b вычисляется как обычно; если же он имеет значение **ЛОЖЬ** или 0, то коэффициент b полагается равным 0 и значения коэффициентов уравнения регрессии вычисляются с учетом этого условия.

Эту функцию можно использовать для одновременного вычисления массива значений $\{\hat{y}\}$ по заданному массиву значений $\{x\}$, для чего в качестве аргумента x надо указать массив $\{x\}$, а саму функцию применить как формулу массива (нажав комбинацию клавиш <Ctrl+Shift+Enter>) к выделенному диапазону ячеек, в котором будет записан выходной массив значений $\{\hat{y}\}$.

4.9.6. Функция ЛГРФПРИБЛ

Применяя метод наименьших квадратов, данная функция рассчитывает коэффициенты экспоненциальной регрессии, т.е. по исходным данным строит функции вида $\hat{Y} = b_0 m^X$ (если исследуется зависимость переменной Y только от одной переменной X) и $\hat{Y} = b_0 \cdot m_1^{x_1} \cdot m_2^{x_2} \cdot \dots \cdot m_k^{x_k}$ (если переменная Y зависит от k переменных X_1, X_2, \dots, X_k). Вид экспоненциальной регрессии, коэффициенты которой m_i и b_0 вычисляет функция **ЛГРФПРИБЛ**, определяется структурой массива значений переменной X : одномерный массив $\{x_1, x_2, \dots, x_n\}$ для регрессии первого вида, двумерный массив $\{x_{11}, x_{12}, \dots, x_{1n}, x_{21}, x_{22}, \dots, x_{2n}, \dots, x_{k1}, x_{k2}, \dots, x_{kn}\}$ для регрессии второго вида. Одним из аргументов функции является также массив $\{y_1, y_2, \dots, y_n\}$ со значениями переменной Y . Функция **ЛГРФПРИБЛ** по массивам исходных данных вычисляет коэффициенты m_i и b_0 , а также может вычислить некоторые статистические характеристики этих коэффициентов и всего уравнения регрессии в целом.

Синтаксис функции:

ЛГРФПРИБЛ(Значения_Y;Значения_X;Константа;Статистика)

Аргумент Значения_Y — одномерный массив (или ссылка на диапазон ячеек, содержащий этот массив) значений Y. Необязательный аргумент Значения_X — массив (или ссылка на диапазон ячеек, содержащий этот массив) значений X. Если данный аргумент опущен, то предполагается, что это массив натуральных чисел {1; 2; 3; ...} такого же размера, как и массив Значения_Y. Если массив Значения_Y расположен в один столбец, то каждый столбец массива Значения_X интерпретируется как значения отдельной переменной X_i . Аналогично, если массив Значения_Y расположен в одну строку, то каждая строка массива Значения_X интерпретируется как значения отдельной переменной X_i .

Аргумент Константа — логическое значение, которое указывает, должен ли коэффициент b_0 быть равным 1. Если этот аргумент имеет значение ИСТИНА, 1 или опущен, то коэффициент b_0 вычисляется как обычно. Если аргумент имеет значение ЛОЖЬ или 0, то b_0 полагается равным 1 и значения коэффициентов m_i вычисляются с учетом этого условия.

Аргумент Статистика принимает логическое значение, которое указывает, требуется ли рассчитывать дополнительные статистические характеристики регрессии. Если он имеет значение ИСТИНА или 1, то функция рассчитывает и выводит эти дополнительные характеристики (см. таблицу в описании функции ЛИНЕЙН; описание и пояснения к этим характеристикам даны в разделе 3.4.3). Если аргумент Статистика имеет значение ЛОЖЬ, 0 или опущен, то функция возвращает только значения коэффициентов m_i и b_0 .

Отметим, что функция возвращает массив значений коэффициентов m_i и b_0 (не менее двух значений), а также дополнительные статистические характеристики (если аргумент Статистика равен ИСТИНА). Поэтому функция должна задаваться в виде формулы массива, в противном случае (при вводе функции в одну ячейку) будет выведено значение только коэффициента m_k . В выходном массиве данные располагаются так же, как и в выходном массиве функции ЛИНЕЙН (см. раздел 4.9.1).

4.9.7. Функция РОСТ

Эта функция является аналогом функции ТЕНДЕНЦИЯ для экспоненциальной регрессии. Она вычисляет в соответствии с построенным методом наименьших квадратов уравнением регрессии значение \hat{Y} для конкретного значения X. Но в отличие от функции ТЕНДЕНЦИЯ эта функция работает с экспоненциальной регрессией.

Пусть задан массив $\{y_1, y_2, \dots, y_n\}$ со значениями переменной Y. Если массив значений переменной X является двумерным массивом вида $\{x_{11}, x_{12}, \dots, x_{1n}, x_{21}, x_{22}, \dots, x_{2n}, \dots, x_{k1}, x_{k2}, \dots, x_{kn}\}$, то в этом случае исследуется зависимость переменной Y от k переменных X_1, X_2, \dots, X_k и строится экспоненциальная регрессия вида $\hat{Y} = b_0 \cdot m_1^{x_1} \cdot m_2^{x_2} \cdot \dots \cdot m_k^{x_k}$. Функция РОСТ по заданным значениям (x_1, x_2, \dots, x_k) и по уравнению регрессии вычисляет значение \hat{y} . Если исходный массив значений X совпадает по размеру с массивом значений Y, то функция РОСТ для вычисления нового значения \hat{y} использует экспоненциальную регрессию вида $\hat{Y} = b_0 m^X$.

Синтаксис функции:

РОСТ(Значения_Y;Значения_X;Новые_значения_x;Константа)

Аргумент Значения_Y — одномерный массив значений Y (или ссылка на диапазон ячеек, содержащий этот массив). Аргумент Значения_X — массив значений X (или ссылка на диапазон ячеек, содержащий этот массив). Аргумент Новые_значения_X — значения, для которых вычисляется уравнение регрессии. Если аргумент Значения_X опущен, то предполагается, что это массив натуральных чисел {1; 2; 3; ...} такого же размера, как и массив аргумента Значения_Y. Если опущен аргумент Новые_значения_X, то по умолчанию предполагается, что он совпадает с аргументом Значения_X.

Аргумент Константа принимает логическое значение: если он имеет значение ИСТИНА или 1 либо опущен, то коэффициент уравнения регрессии b_0 вычисляется как обычно; если же он имеет значение ЛОЖЬ или 0, то коэффициент b_0 полагается равным 1 и значения коэффициентов уравнения регрессии вычисляются с учетом этого условия.

Эту функцию можно использовать для одновременного вычисления массива значений $\{\hat{y}\}$ по заданному массиву значений $\{x\}$, для чего в качестве аргумента Новые_значения_X надо указать массив $\{x\}$, а саму функцию применить как формулу массива к выделенному диапазону ячеек, в котором будет записан выходной массив значений $\{\hat{y}\}$.

4.10. Функции для вычисления ковариации и коэффициента корреляции

В эту группу входят следующие функции.

Функция	Назначение
КОВАР	Возвращает ковариацию, т.е. среднее произведений отклонений для каждой пары точек данных
КОРРЕЛ	Возвращает коэффициент корреляции между двумя наборами данных
ПИРСОН	Возвращает коэффициент корреляции Пирсона
КВПИРСОН	Возвращает квадрат коэффициента корреляции Пирсона
ФИШЕР	Возвращает преобразование Фишера
ФИШЕРОВР	Возвращает функцию, обратную преобразованию Фишера

4.10.1. Функция КОВАР

Функция по двумерной выборке (парным наблюдениям) вычисляет выборочную ковариацию, которая является оценкой ковариации двумерного распределения случайного вектора (X, Y) . Напомним (см. раздел 1.3), что ковариация определяется как математическое ожидание от произведения $(X - MX)(Y - MY)$, т.е. $\text{cov}(X, Y) = M[(X - MX)(Y - MY)]$.

Если имеются парные наблюдения $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ — значения двумерной выборки объемом n , то выборочная ковариация вычисляется по

формуле $\overline{\text{cov}}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$, где $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ и $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

Синтаксис функции:

КОВАР(Массив1;Массив2)

Аргумент Массив1 — первый массив данных (значений переменной X или значений переменной Y; поскольку формула для вычисления ковариации симметрична относительно значений X и Y, не существенно, массив значений какой переменной определять первым). Аргумент Массив2 — второй массив данных.

Аргументы должны быть числами или именами диапазонов, массивами или ссылками на диапазоны ячеек. Если среди значений имеются текстовые или логические значения либо пустые ячейки, то такие значения игнорируются; однако ячейки, которые содержат нулевые значения, учитываются.

Оба аргумента должны содержать одинаковые количества значений. Если они имеют различные объемы данных, то функция возвращает значение ошибки #Н/Д. Если хотя бы один аргумент не задан, то функция возвращает значение ошибки #ДЕЛ/0!

4.10.2. Функция КОРРЕЛ

Данная функция вычисляет выборочный коэффициент корреляции r , т.е.

оценку коэффициента корреляции ρ случайных величин X и Y: $\rho = \frac{\text{cov}(X, Y)}{\sqrt{DX \cdot DY}}$

(см. раздел 1.3). Выборочный коэффициент корреляции r вычисляется по формуле

$r = \frac{\overline{\text{cov}(X, Y)}}{S_x S_y}$, где $\overline{\text{cov}(X, Y)}$ — выборочная ковариация (см. функцию

КОВАР), $S_x = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$, $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ и $S_y = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \bar{y})^2}$, $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

Синтаксис функции:

КОРРЕЛ(Массив1;Массив2)

Аргумент Массив1 — первый массив данных (значений переменной X или значений переменной Y; поскольку формула вычисления коэффициента корреляции симметрична относительно значений X и Y, не существенно, массив значений какой переменной определять первым). Аргумент Массив2 — второй массив данных.

Аргументы должны быть числами или именами диапазонов, массивами или ссылками на диапазоны ячеек. Если среди значений имеются текстовые или логические значения либо пустые ячейки, то такие значения игнорируются; однако ячейки, которые содержат нулевые значения, учитываются.

Оба аргумента должны содержать одинаковые количества значений. Если они имеют различные объемы данных, то функция возвращает значение ошибки #Н/Д. Если хотя бы один аргумент не задан, то функция возвращает значение ошибки #ДЕЛ/0!

4.10.3. Функция ПИРСОН

Эта функция, как и функция КОРРЕЛ, вычисляет выборочный коэффициент корреляции, причем результаты вычислений обеих функций совпадают (конечно, на одном и том же наборе данных). Но в таком случае выборочный

коэффициент корреляции называется *коэффициентом корреляции Пирсона*². Исторически так сложилось, что “обычный” выборочный коэффициент корреляции вычисляется по формулам, приведенным в описании функций КОВАР и КОРРЕЛ. Коэффициент корреляции Пирсона вычисляется по аналогичным формулам, но с использованием несмещенных оценок дисперсий $s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$

и $s_y^2 = \frac{1}{n-1} \sum_{i=1}^n (y_i - \bar{y})^2$; в этом случае оценка ковариации вычисляется по форму-

ле $\text{COV}(X, Y) = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})$. Очевидно, что результаты вычислений в обо-

их случаях будут одинаковы, но “традиция — есть традиция”.

Синтаксис функции:

ПИРСОН(Массив1;Массив2)

Аргумент Массив1 — первый массив данных (значений переменной X или значений переменной Y; поскольку формула вычисления коэффициента корреляции симметрична относительно значений X и Y, не существенно, массив значений какой переменной определять первым). Аргумент Массив2 — второй массив данных.

Аргументы должны быть числами или именами диапазонов, массивами или ссылками на диапазоны ячеек. Если среди значений имеются текстовые или логические значения либо пустые ячейки, то такие значения игнорируются; однако ячейки, которые содержат нулевые значения, учитываются.

Оба аргумента должны содержать одинаковые количества значений. Если они имеют различные объемы данных, то функция возвращает значение ошибки #Н/Д. Если хотя бы один аргумент не задан, то функция возвращает значение ошибки #ДЕЛ/0!.

4.10.4. Функция КВПИРСОН

Функция вычисляет квадрат коэффициента корреляции Пирсона. Эта величина носит название коэффициента детерминации R^2 и показывает при построении линейной регрессии степень точности аппроксимации выборочных данных прямой линией (см. описание функции ЛИНЕЙН и раздел 3.4.3).

Синтаксис функции:

КВПИРСОН(Массив1;Массив2)

Аргумент Массив1 — первый массив данных (значений переменной X или значений переменной Y; поскольку формула вычисления коэффициента корреляции симметрична относительно значений X и Y, не существенно, массив значений какой переменной определять первым). Аргумент Массив2 — второй массив данных³.

Карл Пирсон (Karl Pearson, 1857-1936) — английский математик, биолог, философ, член Лондонского королевского общества, профессор Лондонского университета. Разработал многие методы математической статистики.

Отметим, что в справочной системе Excel при описании этой функции Массив1 обозначается как Значения_Y, а Массив2 — как Значения_X, т.е. предполагается неравнозначность этих аргументов. На самом деле не существенно, какой аргумент представляет значения переменной X, а какой — значения переменной Y. Поэкспериментируйте и убедитесь в этом сами.

Аргументы должны быть числами или именами диапазонов, массивами или ссылками на диапазоны ячеек. Если среди значений имеются текстовые или логические значения либо пустые ячейки, то такие значения игнорируются; однако ячейки, которые содержат нулевые значения, учитываются.

Оба аргумента должны содержать одинаковые количества значений. Если они имеют различные объемы данных, то функция возвращает значение ошибки #Н/Д. Если хотя бы один аргумент не задан, то функция возвращает значение ошибки #ДЕЛ/0!.

4.10.5. Функции ФИШЕР и ФИШЕРОБР

Эти функции обычно используются для построения доверительных интервалов для коэффициентов корреляции или для проверки статистических гипотез о значимости выборочных коэффициентов корреляции (см. раздел 3.2.1). Именно по этой причине данные функции включены в одну группу функций, но, конечно, их можно использовать и в других ситуациях. Практическое использование этих функций показано в главе 13.

Функция ФИШЕР по заданному аргументу x вычисляет значение $Z = \frac{1}{2} \ln \frac{1+x}{1-x}$.

Функция ФИШЕРОБР обратная к функции ФИШЕР. Она по заданному аргументу x вычисляет значение $y = \frac{e^{2x} - 1}{e^{2x} + 1}$.

Синтаксис функций:

ФУНКЦИЯ(x)

Если аргумент x не является числом, то функции возвращают значение ошибки #ЗНАЧ!. Аргумент x функции ФИШЕР должен быть из интервала $(-1, 1)$, иначе функция возвращает значение ошибки #ЧИСЛО!.

4.11. Дополнительные функции

В эту группу вошли функции, которые было трудно поместить в какую-либо из приведенных выше групп функций. Поэтому они несколько "разношерстные", но все как-то связаны с вычислением вероятностей. В следующем разделе собраны все остальные функции, которые еще остались в категории статистических функций.

Функция	Назначение
ВЕРОЯТНОСТЬ	Возвращает вероятность того, что заданные значения находятся внутри определенного интервала
ДОВЕРИТ	Возвращает доверительный интервал для среднего генеральной совокупности *
МОДА	Возвращает моду (наиболее часто встречающееся значение) набора данных
ЧАСТОТА	Возвращает массив чисел, равных количеству выборочных значений, попадающих в заданное множество интервалов

4.11.1. Функция ВЕРОЯТНОСТЬ

Эта функция работает с произвольным дискретным распределением. Для ее работы необходимо иметь массив значений x_i и массив вероятностей p_i , с которыми принимаются эти значения, т.е. вероятностную таблицу следующего вида.

Значения x_i	x_1	x_2	...	x_n
Вероятности p_i	p_1	p_2	...	p_n

Функция не требует, чтобы значения x_i были отсортированы в порядке возрастания или не имели одинаковых значений. Необходимо только, чтобы сумма всех вероятностей p_i равнялась 1.

Функция ВЕРОЯТНОСТЬ по заданному значению x определяет вероятность этого значения; если среди значений x_i нет значения, совпадающего с заданным значением x , то функция возвращает значение 0. Можно также задать интервал $[a, b]$ — и функция определит, сколько значений x_i попадает в этот интервал и вернет вероятность, равную сумме вероятностей тех x_i , которые попадают в этот интервал.

Синтаксис функции:

ВЕРОЯТНОСТЬ(Интервал_X;Интервал_P;a;b)

Аргумент Интервал_X — интервал числовых значений x_i . Аргумент Интервал_P — интервал вероятностей, соответствующих значениям в аргументе Интервал_X. Аргумент a — нижняя граница интервала $[a, b]$, для которого вычисляется вероятность. Необязательный аргумент b — верхняя граница интервала $[a, b]$. Если аргумент b не задан, то аргумент a считается тем значением x , для которого находится вероятность.

Если любое значение в аргументе Интервал_P меньше 0 или больше 1, то функция возвращает значение ошибки #ЧИСЛО!. Такое же значение ошибки функция возвращает в случае, если сумма значений в аргументе Интервал_P не равна 1.

Если Интервал_X и Интервал_P содержат различные количества значений, то функция возвращает значение ошибки #Н/Д.

4.11.2. Функция ДОВЕРИТ

Данная функция используется при построении доверительного интервала для неизвестного математического ожидания генеральной совокупности, имеющей нормальное распределение, при условии, что дисперсия σ^2 этого распределения известна.

Напомним (см. раздел 2.3.6), что для точечного оценивания математического ожидания m используется статистика $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$, а доверительный интервал для m

с доверительным уровнем p определяется как $\left(\bar{x} - k \frac{\sigma}{\sqrt{n}}, \bar{x} + k \frac{\sigma}{\sqrt{n}} \right)$, где n — объем

выборки, значение k находится по формуле $k = \Phi^{-1}\left(\frac{1+p}{2}\right)$, Φ^{-1} — функция, обратная к функции распределения стандартного нормального закона. Функция ДОВЕРИТ по заданным значениям $\alpha = 1 - p$, σ и n вычисляет величину $k \frac{\sigma}{\sqrt{n}}$.

Синтаксис функции:

ДОВЕРИТ(Альфа;Станд_отклонение;Размер)

Аргумент **Альфа** — уровень значимости, связанный с доверительным уровнем p соотношением $p = 1 - \text{Альфа}$. Аргумент **Станд_отклонение** — известное стандартное отклонение σ генеральной совокупности. **Размер** — объем выборки n . Если значение этого аргумента — нецелое число, то берется целая часть этого числа.

Если какой-либо из аргументов не является числом, то функция возвращает значение ошибки #ЗНАЧ!. Если **Альфа** меньше 0 или больше 1, то функция возвращает значение ошибки #ЧИСЛО!. Это же значение функция возвращает в случае, если значение **Станд_отклонение** отрицательно или значение аргумента **Размер** меньше 1.

4.11.3. Функция МОДА

Данная функция имеет мало общего с нахождением моды, т.е. наибольшего значения плотности вероятности непрерывного распределения (см. раздел 1.2.3). Поэтому она не включена в группу функций, вычисляющих геометрические характеристики распределения. Функция **МОДА** среди заданного набора числовых значений $\{x_1, x_2, \dots, x_n\}$ находит значение, которое повторяется наиболее часто. Если одинаковых значений нет, то функция возвращает значение ошибки #Н/Д. Если несколько значений повторяются одно и то же количество раз, то выводится первое такое значение.

Синтаксис функции:

МОДА(Число1;Число2;...)

Функция может иметь до 30 аргументов **Число**. Эти аргументы могут быть числами, именами диапазонов, массивами или ссылками на диапазоны. Можно использовать один массив или одну ссылку на диапазон вместо аргументов, разделяемых точкой с запятой.

Если аргумент, являясь массивом или ссылкой, содержит текст, логические значения или пустые ячейки, эти значения игнорируются; ячейки, содержащие нулевые значения, учитываются.

4.11.4. Функция ЧАСТОТА

Эта функция часто используется для построения гистограмм (см. раздел 8.3). Функция подсчитывает, сколько значений из заданного массива значений $\{x_1, x_2, \dots, x_n\}$ попадает в интервалы $(-\infty, a_1]$, $(a_1, a_2]$, ..., $(a_{k-1}, a_k]$, $(a_k, +\infty)$ (такие интервалы часто называют *карманами*). Интервалы задаются набором чисел $\{a_1, a_2, \dots, a_{k-1}, a_k\}$. Хотя функция этого не требует, но логично, чтобы выполнялось условие $a_1 < a_2 < \dots < a_{k-1} < a_k$. Итак, функция **ЧАСТОТА** возвращает массив чисел размером $k + 1$. Поэтому она должна применяться как *формула массива* к выделенному диапазону ячеек, состоящему не менее чем из $k + 1$ ячейки. Если ее применить в одной ячейке, то она вернет только количество значений x_i , попавших в интервал $(-\infty, a_1]$.

Синтаксис функции:

ЧАСТОТА(Массив_данных;Массив_интервалов)

Аргумент **Массив_данных** — массив или ссылка на диапазон ячеек, содержащий значения $\{x_1, x_2, \dots, x_n\}$. Если **Массив_данных** не содержит значений, то функция

возвращает массив нулей. Аргумент Массив_интервалов — массив или ссылка на диапазон ячеек, содержащий значения границ интервалов $\{a_1, a_2, \dots, a_{k-1}, a_k\}$. Функция игнорирует пустые ячейки, а также текстовые и логические значения.

4.12. Вспомогательные функции

Рассмотрим последние функции категории Статистические, которые выполняют вспомогательные вычисления.

Функция	Назначение
ГАММАНЛОГ	Возвращает натуральный логарифм гамма-функции
НОРМАЛИЗАЦИЯ	Возвращает нормализованную величину
ПЕРЕСТ	Возвращает число перестановок для заданного числа объектов
СЧЁТ	Подсчитывает количество чисел в списке аргументов
СЧЁТЗ	Подсчитывает количество непустых значений в списке аргументов

4.12.1. Функция ГАММАНЛОГ

Значения гамма-функции Эйлера $\Gamma(x) = \int_0^\infty e^{-u} u^{x-1} du$ часто используются в статистических расчетах, поскольку она участвует в формулах вычисления плотности вероятности и функции распределения многих распределений, например распределения Стюдента, распределения χ^2 , F-распределения и др. Поскольку значение функции $\Gamma(x)$ быстро растет при возрастании значения x (например, если x — натуральное число, $\Gamma(x) = (x-1)!$), то на практике удобнее использовать логарифм от этой функции. Функция ГАММАНЛОГ и вычисляет натуральный логарифм от функции $\Gamma(x)$. Чтобы получить значение самой функции, следует применить формулу

$$= \text{EXP}(\text{ГАММАНЛОГ}(x))$$

Синтаксис функции:

$$\text{ГАММАНЛОГ}(x)$$

Аргумент x — это значение, для которого вычисляется функция. Если x не является числом, то функция возвращает значение ошибки #ЗНАЧ!. Если x меньше или равно 0, то функция возвращает значение ошибки #ЧИСЛО!.

4.12.2. Функция НОРМАЛИЗАЦИЯ

Если случайная величина X имеет математическое ожидание μ и дисперсию σ^2 , то случайная величина $Y = (X - \mu)/\sigma$ имеет то же распределение, что и случайная величина X , но с математическим ожиданием 0 и дисперсией $\sigma^2 = 1$. Такая операция, преобразование $Y = (X - \mu)/\sigma$, называется *нормализацией* случайной величины X . Данная функция и выполняет такую операцию.

Синтаксис функции:

$$\text{НОРМАЛИЗАЦИЯ}(x; \text{Среднее}; \text{Стандартное_отклонение})$$

Аргумент *x* — нормализуемое значение. Аргумент Среднее — задаваемое математическое ожидание μ . Аргумент Стандартное_отклонение — среднеквадратическое отклонение σ .

Если Стандартное_отклонение меньше или равно 0, то функция возвращает значение ошибки #ЧИСЛО!.

4.12.3. Функция ПЕРЕСТ

Функция вычисляет количество перестановок для заданного числа *k* объектов, которые выбираются из общего числа *n* объектов. Перестановка — это любое множество или подмножество объектов, которые отличаются либо составом объектов, либо их порядком. Перестановки отличаются от сочетаний, для которых внутренний порядок не имеет значения. Эта функция используется для вычисления вероятностей в комбинаторных задачах. Число перестановок обычно обозначается как $P_{k,n}$ и вычисляется по формуле

$$P_{k,n} = \frac{n!}{(n-k)!} = n(n-1)(n-2)\dots(n-k+1)$$

Синтаксис функции:

ПЕРЕСТ(Число;Число_выбранных)

Аргумент Число — целое число *n*, задающее количество объектов. Аргумент Число_выбранных — целое число *k*, задающее количество объектов в каждой перестановке.

Если аргументы не являются целыми числами, то берется целая часть этих чисел. Если аргументы не являются числами, то функция возвращает значение ошибки #ЗНАЧ!. Если значения аргументов отрицательны, то функция возвращает значение ошибки #ЧИСЛО!. Это же значение ошибки функция возвращает в том случае, если значение аргумента Число меньше значения аргумента Число_выбранных.

4.12.4. Функции СЧЁТ и СЧЁТЗ

Эти функции подсчитывают количество чисел (функция СЧЁТ) и количество непустых ячеек (функция СЧЁТЗ) в заданном диапазоне ячеек.

Синтаксис функций:

ФУНКЦИЯ(Значение1;Значение2;...)

Функции могут иметь до 30 аргументов Значение, которые могут быть значениями, массивами, именами или адресами диапазонов.

Отметим, что в Excel имеются функции СЧЁТЕСЛИ и СЧИТАТЬПУСТОТЫ, которые также можно использовать для подсчета количества значений.

4.13. Функции для генерирования равномерно распределенных случайных чисел

Эти функции не входят в категорию Статистические (они входят в категорию Математические), часто используются в статистическом анализе для моделирования случайных величин (см. главу 7) и без них описание статистических функций Excel было бы не полным.

Функция	Назначение
СЛЧИС	Генерирует равномерно распределенные на интервале [0, 1] случайные числа
СЛУЧМЕЖДУ	Генерирует целые числа, равномерно распределенные на заданном интервале

4.13.1. Функция СЛЧИС

Эта функция возвращает числа, равномерно распределенные на интервале [0, 1]. Ее синтаксис — СЛЧИС(), т.е. она не имеет аргументов. Она часто используется для генерирования случайных чисел методом обратной функции (см. главу 7), а также в имитационном моделировании.

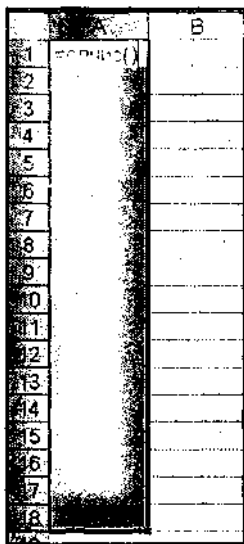


Рис. 4.9. Создание диапазона со случайными числами

Функцию СЛЧИС можно использовать в формулах массивов для генерирования диапазонов случайных чисел. На рис. 4.9 показан процесс создания диапазона случайных чисел. Сначала выделяется диапазон ячеек, затем, не снимая выделения, вводится формула =СЛЧИС() и после этого нажимается комбинация клавиш <Ctrl+Shift+Enter>.

Необходимо отметить, что формулы, содержащие функцию СЛЧИС, пересчитываются при каждом пересчете рабочего листа, например при вводе любого значения в ячейку или при удалении чего-либо. Это свойство данной функции полезно, например, в имитационном моделировании. Однако в других случаях оно может сильно замедлять работу в Excel или быть просто излишним. Чтобы зафиксировать значения, вычисляемые с помощью функции СЛЧИС, надо выделить диапазон ячеек, содержащий эти значения, и скопировать его (команда Правка^Копировать). Затем, не снимая выделения диапазона, следует выполнить команду Правка^Специальная вставка, в открывшемся диалоговом окне Специальная вставка установить переключатель Значения, как показано на рис. 4.10, и щелкнуть на кнопке ОК. В ячейки выделенного диапазона вместо формул будут записаны числовые значения.

Применение функции СЛЧИС для генерирования случайных чисел, которые имеют распределения, отличные от равномерного, показано в главе 7.

4.13.2. Функция СЛУЧМЕЖДУ

Эта функция генерирует целочисленные значения, подчиняющиеся дискретному равномерному распределению (см. раздел 1.4.1). Отметим, что она доступна только тогда, когда подключена надстройка Пакет анализа.

Синтаксис функции:

СЛУЧМЕЖДУ(Нижняя_граница;Верхняя_граница)

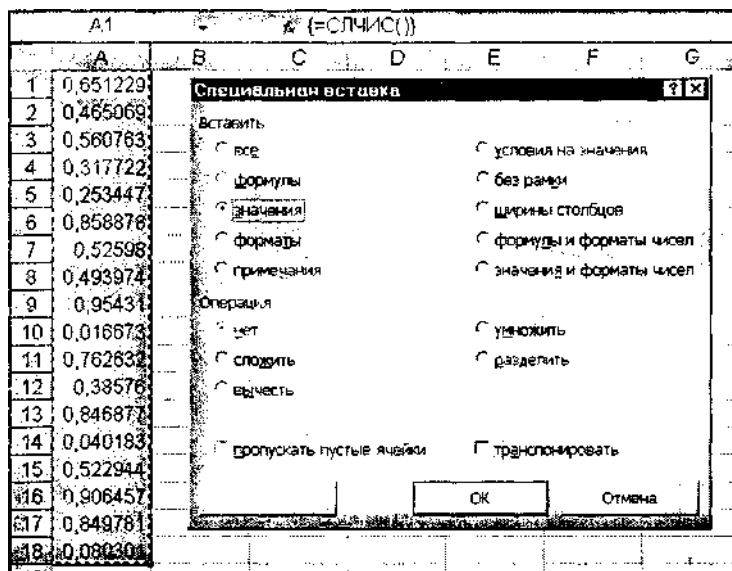


Рис. 4.10. Преобразование формул в числовые значения

Аргумент *Нижняя_граница* задает нижнюю границу интервала изменения случайной величины, аргумент *Верхняя_граница* — верхнюю границу этого интервала. Если значения аргументов дробные, они округляются до ближайших целых. Если значение аргумента *Нижняя_граница* больше значения аргумента *Верхняя_граница*, функция возвращает значение ошибки #ЧИСЛО!.

Эта функция "не работает" в формулах массивов. Поэтому, чтобы сгенерировать диапазон значений, сначала надо ввести формулу =СЛУЧМЕЖДУ(-10;10) в первую ячейку, а затем скопировать ее во все остальные ячейки диапазона. (Приведенная формула будет генерировать целые числа из интервала от -10 до 10.)

Как и в случае с функцией СЛЧИС, формулы, содержащие функцию СЛУЧМЕЖДУ, пересчитываются при каждом пересчете рабочего листа. Поэтому, чтобы зафиксировать значения, полученные с помощью этой функции, следует преобразовать формулы в значения, как описано в предыдущем разделе.

Надстройка Пакет анализ

В состав Microsoft Excel входит надстройка Пакет анализа, которая содержит 19 статистических процедур и около 50 функций. Функции в основном относятся к категориям инженерных и финансовых и поэтому здесь не рассматриваются. Статистические процедуры, содержащиеся в надстройке Пакет анализа, предоставляют широкий спектр средств для статистического анализа начиная от простой описательной статистики или сглаживания данных и заканчивая анализом Фурье и проведением различных тестов. Полный список этих средств и их краткое описание представлены в табл. 5.1 (названия средств приводятся в соответствии со списком из диалогового окна Анализ данных).

Таблица 5.1. Статистические средства надстройки Пакет анализа

<i>Средство</i>	<i>Описание</i>
Однофакторный дисперсионный анализ	Используется для проверки гипотезы о равенстве математических ожиданий двух или более выборок
Двухфакторный дисперсионный анализ без повторений	Двухфакторный дисперсионный анализ на основе одной выборки
Двухфакторный дисперсионный анализ с повторениями	Двухфакторный дисперсионный анализ на основе нескольких выборок
Корреляция	Вычисляет корреляционную матрицу
Ковариация	Вычисляет матрицу ковариаций
Описательная статистика	Создает отчет, содержащий статистические характеристики представленной выборки
Экспоненциальное сглаживание	Реализует метод экспоненциального сглаживания данных
Двухвыборочный F-тест для дисперсий	Применяется для сравнения дисперсий двух генеральных совокупностей
Анализ Фурье-	Реализует метод быстрого преобразования Фурье (БПФ) для анализа данных
Гистограмма	Используется для анализа распределения выборочных данных и построения гистограмм
Скользящее среднее	Используется для сглаживания данных
Генерация случайных чисел	Генерирует случайные числа, имеющие заданное распределение
Ранг и персентиль	Используется для вычисления рангов и квантилей

<i>Средство</i>	<i>Описание</i>
Регрессия	Используется для построения линейной регрессии
Выборка	Создает случайную выборку, рассматривая входной диапазон значений как генеральную совокупность
Парный двухвыборочный t-тест для средних	Используется для проверки гипотезы о равенстве математических ожиданий для двумерной выборки данных
Двухвыборочный t-тест с одинаковыми дисперсиями	Служит для проверки гипотезы о равенстве математических ожиданий для двух выборок. Предполагается равенство дисперсий генеральных совокупностей
Двухвыборочный t-тест с разными дисперсиями	Используется для проверки гипотезы о равенстве математических ожиданий для двух выборок. Не требует предположения о равенстве дисперсий генеральных совокупностей
Двухвыборочный z-тест для средних	Используется для проверки гипотезы о различии между математическими ожиданиям двух генеральных совокупностей

Отметим, что эти средства имеют определенные ограничения и иногда удобнее воспользоваться статистическими функциями или другими средствами Excel. Преимуществом функций перед данными средствами является то, что функции автоматически пересчитываются при любых изменениях, сделанных в выборке, тогда как эти средства необходимо выполнять заново, если выборка изменилась. В "оправдание" этих средств скажем, что они сохраняют установки, сделанные пользователем при последнем применении средства, но только в течение одного сеанса работы с Excel.

Средства, которые включены в надстройку Пакет анализа, доступны через команду Сервис^Анализ данных. (Если команды Анализ данных нет в меню Сервис, подключите эту надстройку. Для этого выполните команду Сервисе Надстройки и в открывшемся диалоговом окне Надстройки в списке Доступные надстройки установите флажок напротив опции Пакет анализа.) Команда Сервис^Анализ данных открывает одноименное диалоговое окно, в списке Инструменты анализа которого следует выбрать необходимое средство (рис. 5.1). После выбора какого-либо средства (и последующего щелчка на кнопке ОК) открывается диалоговое окно этого средства.

В большинстве таких диалоговых окон (на рис. 5.2 для примера показано диалоговое окно средства Описательная статистика) выделены области Входные данные и Параметры вывода. В области Входные данные указывается диапазон ячеек, в котором содержатся данные (поле Входной интервал), указывается, сгруппированы ли данные, и если сгруппированы, то по столбцам или по строкам (переключатели по столбцам и по строкам). Если задается входной диапазон данных вместе с заголовками, то устанавливается флажок опции Метки в первой строке (столбце). (Если заголовки не задаются, то данным автоматически присваиваются заголовки Столбец1, Столбец2 и т.д. или Строка1, Строка2 и т.д. в зависимости от того, расположены данные в столбцах или в строках.) В некоторых

диалоговых окнах в области Входные данные необходимо указать несколько входных диапазонов (например, в окне Регрессия) либо дополнительные параметры для проведения выбранной статистической процедуры, например доверительный уровень для проведения тестов.

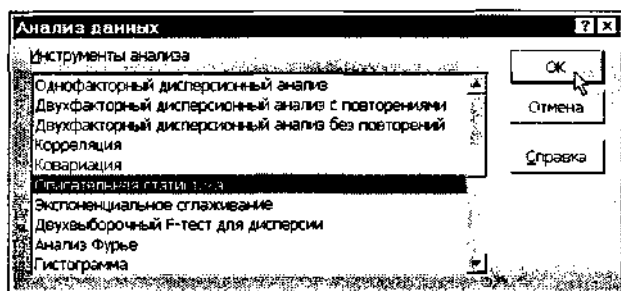


Рис. 5.1. Диалоговое окно Анализ данных со списком инструментов статистического анализа

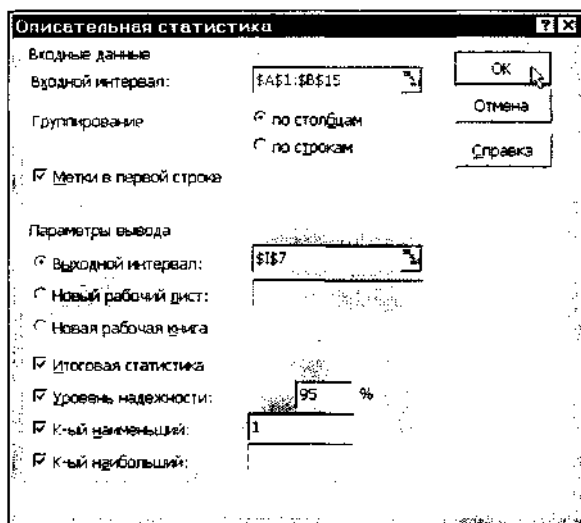


Рис. 5.2. Диалоговое окно средства Описательная статистика

В области Параметры вывода, как правило, надо указать, куда будут выводиться результаты расчетов. Предусмотрено три возможности: на текущий рабочий лист (переключатель Выходной интервал), при этом необходимо указать выходной интервал (достаточно указать адрес одной ячейки, которая определяет верхний левый угол выходного диапазона); на новый рабочий лист текущей рабочей книги начиная с ячейки A1 (переключатель Новый рабочий лист), при этом можно сразу задать имя этому листу; в новую рабочую книгу (переключатель Новая рабочая книга), в этом случае автоматически открывается новая рабочая книга. Также в этой области часто имеются опции, которые указывают, что

именно необходимо вывести из возможного набора выходных результатов (например, графики либо дополнительные статистические характеристики).

В некоторых диалоговых окнах имеются другие области, в которых содержатся опции, необходимые для работы данного средства. Эти опции будут приведены при описании конкретных средств. Опции областей Входные данные и Параметры вывода будем упоминать только тогда, когда они будут отличаться от описанных выше.

Перейдем к описанию конкретных средств статистического анализа, при этом будем называть их так, как они названы в списке диалогового окна Анализ данных. Опишем их в порядке "от простого к сложному" (другими словами, в том порядке, который больше нравится автору).

5.1. Описательная статистика

Это средство (вместе со средством Гистограмма, которое будет описано в следующем разделе) является, по-видимому, наиболее часто используемым из всего пакета анализа, поскольку быстро и просто вычисляет основные статистические характеристики одномерных выборок. На рис. 5.3 показан рабочий лист, содержащий три ряда данных (три независимые выборки, имеющие разные распределения) и диалоговое окно Описательная статистика.

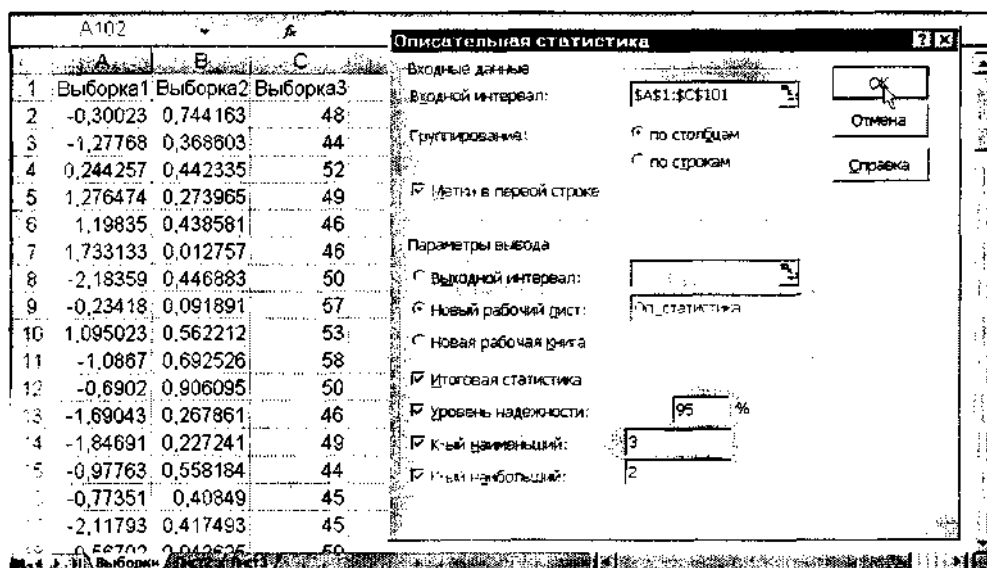


Рис. 5.3. Три выборки и диалоговое окно Описательная статистика

Отметим, что в данном случае имеются выборки разных размеров. Средство "описательная статистика" правильно определяет размеры выборок, игнорируя пустые ячейки. На рис. 5.4 показан рабочий лист с результатами расчетов.

В табл. 5.2 перечислены вычисляемые средством Описательная статистика статистические характеристики выборок, а также функции, которые возвращают те же самые характеристики.

	A	B	C	D	E	F
1	Выборка1		Выборка2		Выборка3	
2						
3	Среднее	-0,04048	Среднее	0,483938	Среднее	48,8571
4	Стандартная ошибка	0,10856	Стандартная ошибка	0,035805	Стандартная ошибка	0,81822
5	Медиана	-0,0848	Медиана	0,444609	Медиана	49
6	Мода	#N/D	Мода	#N/D	Мода	50
7	Стандартное отклонение	1,08583	Стандартное отклонение	0,253178	Стандартное отклонение	3,68335
8	Дисперсия выборки	1,17859	Дисперсия выборки	0,064099	Дисперсия выборки	13,4202
9	Экссесс	-0,47571	Экссесс	-0,53969	Экссесс	0,81597
10	Асимметричность	0,0907	Асимметричность	0,278514	Асимметричность	0,75294
11	Интервал	4,95324	Интервал	0,961028	Интервал	15
12	Минимум	-2,57758	Минимум	0,012757	Минимум	43
13	Максимум	2,37565	Максимум	0,973785	Максимум	58
14	Сумма	-4,04848	Сумма	24,19681	Сумма	1710
15	Счет	100	Счет	50	Счет	35
16	Наибольший(3)	2,1945	Наибольший(3)	0,946319	Наибольший(3)	57
17	Наименьший(2)	-2,19359	Наименьший(2)	0,056185	Наименьший(2)	44
18	Уровень надежности(95,0%)	0,21541	Уровень надежности(95,0%)	0,071952	Уровень надежности(95,0%)	1,25841
19						
20						

Рис. 5.4. Результаты работы средства Описательная статистика

Таблица 5.2. Значения, вычисляемые средством Описательная статистика

Значение	Описание
Среднее	Выборочное среднее $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$. Функция СРЗНАЧ
Стандартная ошибка	Оценка среднеквадратического отклонения выборочного среднего; вычисляется по формуле $\sqrt{\frac{1}{n(n-1)} \sum_{i=1}^n (x_i - \bar{x})^2}$
Медиана	Значение медианы, т.е. квантиля порядка 0,5. Функция МЕДИАНА
Мода	Значение моды. Вычисляется так же, как и функцией МОДА (см. раздел 4.11.3), — если нет одинаковых выборочных значений, то возвращается значение ошибки #N/D
Стандартное отклонение	Оценка среднеквадратического отклонения генеральной совокупности $s_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$. Функция СТАНДОТКЛОН
Дисперсия выборки	Оценка дисперсии генеральной совокупности $s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$. Функция ДИСП
Экссесс	Выборочный коэффициент эксцесса (см. раздел 2.3.4). Функция ЭКСЦЕСС
Асимметричность	Выборочный коэффициент асимметрии (см. раздел 2.3.4). Функция СКОС

<i>Значение</i>	<i>Описание</i>
Интервал	Размах выборки. Вычисляется как разность между максимальным и минимальным выборочными значениями
Минимум	Минимальное выборочное значение. Функция МИН
Максимум	Максимальное выборочное значение. Функция МАКС
Сумма	Сумма выборочных значений. Функция СУММ
Счет	Объем выборки. Функция СЧЁТ
Наибольший (K)	K-е наибольшее значение. Если K = 1, то выводится максимальное выборочное значение. Функция НАИБОЛЬШИЙ
Наименьший (K)	K-е наименьшее значение. Если K = 1, то выводится минимальное выборочное значение. Функция НАИМЕНЬШИЙ
Уровень надежности (X%)	Граница доверительного интервала для неизвестного математического ожидания с доверительным уровнем X%; доверительный интервал строится как выборочное среднее плюс-минус данное значение. Граница вычисляется с помощью распределения Стьюдента (см. раздел 2.3.6), т.е. 'здесь неявно используется предположение о нормальности распределения генеральной совокупности. Поэтому к данному показателю следует относиться осторожно, особенно при малых выборках

5.1.1. Опции диалогового окна **Описательная статистика**

Установка флажка опции Итоговая статистика указывает, что в итоговом отчете этого средства будут вычислены все статистические характеристики выборки, за исключением границы доверительного интервала для среднего и K-х наибольших и наименьших значений, для которых имеются отдельные опции Уровень надежности, K-ый наименьший и K-ый наибольший. Если флажок опции Итоговая статистика не установлен, то выводится только то, что задается с помощью опций Уровень надежности, K-ый наименьший и K-ый наибольший.

Опция Уровень надежности указывает, надо ли вычислять границу доверительного интервала для среднего. В поле ввода рядом с этой опцией задается доверительный уровень в процентах.

В полях ввода рядом с опциями K-ый наибольший и K-ый наименьший указываются порядки выводимых наибольшего и наименьшего значений. Если эти порядки равны 1, то выводятся соответственно максимальное и минимальное выборочные значения.

5.2. Гистограмма

Это средство полезно для первичного анализа распределения выборки и построения гистограмм (столбцовых диаграмм эмпирических плотностей вероятностей). В качестве исходных данных нужно указать входной диапазон, содержащий выборочные значения, и интервал карманов. Интервал карманов определяет границы для столбцов гистограммы. Средство Гистограмма подсчитывает число выборочных значений, попавших в каждый карман (эти числа в выходных данных

называются Частота), и по этим числам строит гистограмму. Далее последовательно суммируются частоты (подсчитываются так называемые накапливающие суммы), эти суммы делятся на объем выборки и умножаются на 100. Получается то, что здесь называется Интегральный процент. На самом деле, если убрать проценты (т.е. накапливающие суммы нормировать не на 100%, а на 1), это просто эмпирическая функция распределения. Средство Гистограмма предоставляет возможность вывести значения интегрального процента в виде графика. В качестве дополнительной возможности предусмотрена сортировка частот по убыванию и построение гистограммы по этим отсортированным частотам.

5.2.1. Опции диалогового окна Гистограмма

Диалоговое окно Гистограмма показано на рис. 5.5. В области Входные данные задаются адрес диапазона ячеек с выборочными значениями (поле ввода Входной интервал) и адрес диапазона, содержащего границы карманов (поле ввода Интервал карманов). Границы карманов должны быть представлены в порядке возрастания. При подсчете количества попаданий выборочных значений в карманы в число попавших в данный карман включаются значения, равные нижней границе кармана и меньшие верхней границы кармана. Если не указывать интервал границ карманов, будут автоматически созданы равновеликие интервалы, количество которых определяется по формуле Стерджесса $h = [1 + 3,22 \ln(r)/x]$ ($[x]$ — целая часть числа x). (Более подробно о построении интервалов речь идет в разделе 8.3.2.)

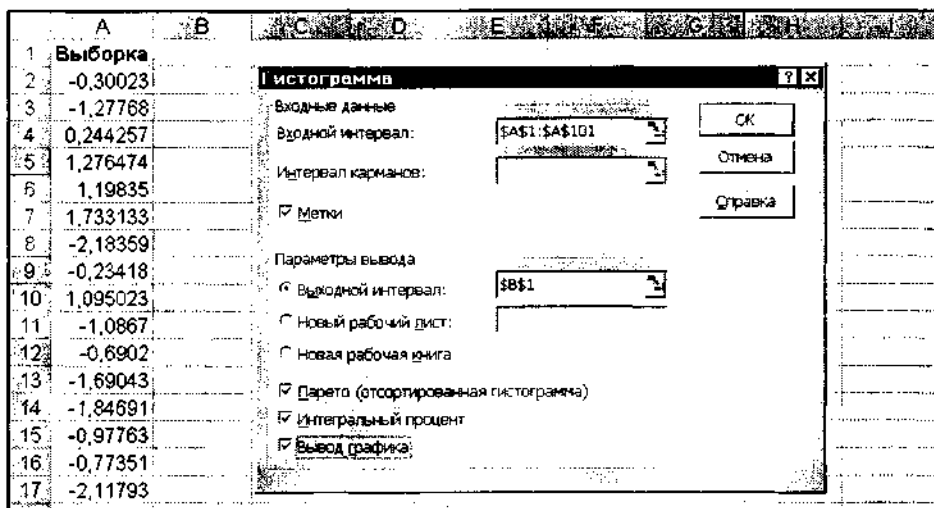


Рис. 5.5. Исходные данные и диалоговое окно Гистограмма

Рассмотрим опции Парето (отсортированная гистограмма), Интегральный процент и Вывод графика из области Параметры вывода.

Если установлен только флажок опции Парето (отсортированная гистограмма), то выводятся таблица частот и таблица отсортированных в порядке убывания частот. Если также установлен флажок опции Вывод графика, выводится гистограмма отсортированных частот, как показано на рис. 5.6.

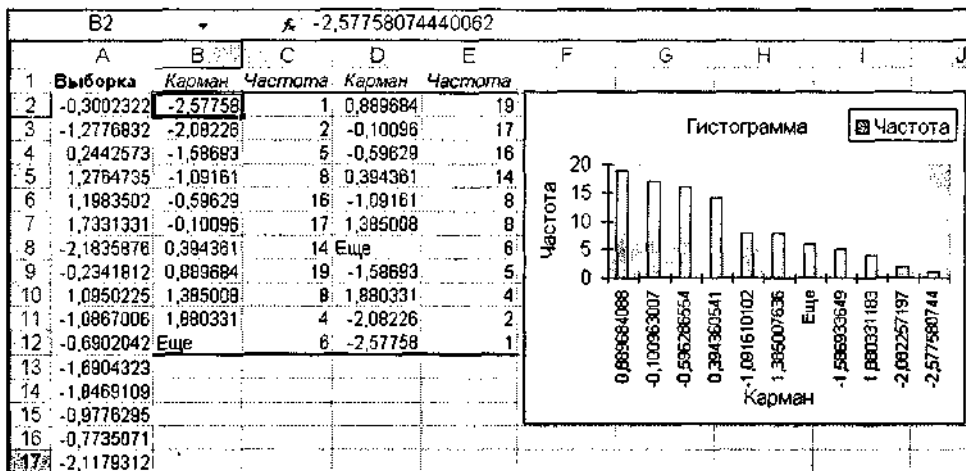


Рис. 5.6. Отсортированная гистограмма

Если установлен только флажок ОПЦИИ Интегральный процент, то выводится таблица, содержащая частоты и значения интегрального процента. Если еще установлен флажок опции Вывод графика, эти данные также отображаются графически, как показано на рис. 5.7.

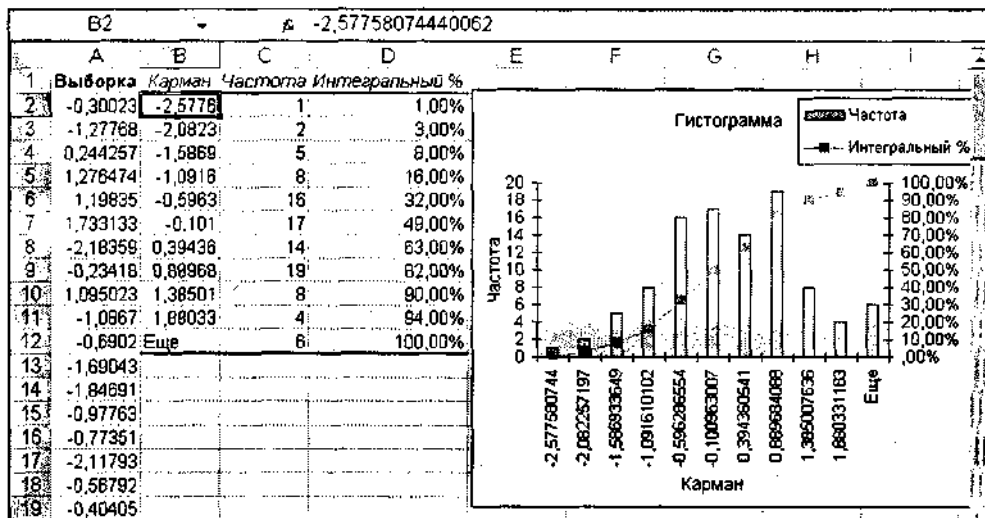


Рис. 5.7. Гистограмма частот и график интегрального процента

Если установлены флажки опций Парето (отсортированная гистограмма) и Интегральный процент, то выводятся две таблицы: одна содержит неотсортированные частоты и интегральные проценты, вторая — отсортированные частоты и соответствующие интегральные проценты (рис. 5.8). Если также установлен флажок опции Вывод графика, выводятся гистограмма и график интегрального процента, построенные по отсортированным частотам.

	A	B	C	D	E	F	G
1	Выборка	Карман	Частота	Интегральный %	Карман	Частота	Интегральный %
2	-0,30023	-2,577581	1	1,00%	0,88968	19	19,00%
3	-1,27768	-2,082257	2	3,00%	-0,101	17	36,00%
4	0,244257	-1,586934	5	8,00%	-0,5963	16	52,00%
5	1,276474	-1,09161	8	16,00%	0,39436	14	66,00%
6	1,19835	-0,596287	16	32,00%	-1,0916	8	74,00%
7	1,733133	-0,100963	17	49,00%	1,38501	8	82,00%
8	-2,18359	0,3943605	14	63,00%	Еще	6	88,00%
9	-0,23418	0,8896841	19	82,00%	-1,5869	5	93,00%
10	1,095023	1,3850076	8	90,00%	1,88033	4	97,00%
11	-1,0867	1,8803312	4	94,00%	-2,0823	2	99,00%
12	-0,6902	Еще	6	100,00%	-2,5776	1	100,00%
13	-1,69043						
14	-1,84691						

Рис. 5.8. Выходные данные (две таблицы)

Наконец, если установлен флажок только опции Вывод графика, выводятся таблица частот (не отсортированная) и гистограмма.

5.3. Генерация случайных чисел

Это средство предназначено для генерирования значений случайных чисел, имеющих заданное распределение, т.е. для получения случайных выборок. Средство имеет возможность генерировать случайные числа, имеющие следующие распределения.

- **Равномерное.** Генерируется последовательность равномерно распределенных случайных чисел в заданном интервале, для чего необходимо указать верхнюю и нижнюю границы интервала.
- **Нормальное.** Генерируется последовательность случайных чисел, подчиняющихся нормальному распределению. Задается математическое ожидание и среднеквадратическое отклонение.
- **Бернулли.** Генерируется последовательность случайных чисел, принимающих только значение 0 или 1, в зависимости от заданной вероятности успеха (исхода "1")- (О распределении Бернулли речь идет в разделе 1.4.2.)
- **Биномиальное.** Генерируется последовательность случайных чисел, равная количеству исходов "1" в n независимых испытаниях. В результате каждого из них с вероятностью p может произойти исход "1" и с вероятностью $(1 - p)$ — исход "0" (см. раздел 1.4.3). Здесь необходимо задать число испытаний n и вероятность p .
- **Пуассона.** Генерируется последовательность случайных чисел, подчиняющихся распределению Пуассона с заданным параметром X . (О распределении Пуассона речь идет в разделе 1.4.4.)
- **Модельное.** При выборе этого распределения на самом деле генерируются не случайные числа, а повторяющаяся последовательность членов арифметической прогрессии, причем члены прогрессии также могут повторяться заданное число раз. Для этого распределения задаются интервал изменения

членов арифметической прогрессии, шаг прогрессии, число повторений членов прогрессии и число повторений этой последовательности чисел.

- Дискретное. Генерируется последовательность случайных чисел, подчиняющихся заданному дискретному распределению. Для задания этого распределения необходимо указать диапазон ячеек, состоящий из двух столбцов: в первом столбце содержатся значения, а во втором — вероятности каждого значения. Сумма вероятностей во втором столбце должна быть равна 1.

5.3.1. Опции диалогового окна Генерация случайных чисел

Диалоговое окно Генерация случайных чисел при задании различных распределений имеет ряд одинаковых элементов, но наличие некоторых других опций зависит от выбранного типа распределения. Выбор распределения осуществляется в раскрывающемся списке Распределение.

Рассмотрим сначала общие элементы всех диалоговых окон Генерация случайных чисел.

В поле ввода Число переменных указывается количество генерируемых выборок. Каждая выборка располагается в отдельном столбце. Максимальное количество выборок — 256 (по количеству столбцов в рабочем листе Excel). Если это число не введено, то будет сгенерирована одна случайная выборка, или, если в поле Выходной интервал указан диапазон ячеек, в котором будут располагаться сгенерированные значения, будут заполнены все столбцы этого диапазона.

В поле ввода Число случайных чисел задается количество выборочных значений (т.е. объем генерируемых выборок), одно и то же для всех выборок. Если это число не введено, то будет сгенерировано одно значение, или, если в поле Выходной интервал указан диапазон ячеек, в котором будут располагаться сгенерированные значения, будут заполнены все строки этого диапазона.

В большинстве диалоговых окон Генерация случайных чисел (кроме окон для модельного и дискретного распределений) имеется поле ввода Случайное рассеивание. Число, введенное в это поле, задает начальное значение, которое будет использовано в алгоритме генерации случайных чисел. Обычно это поле оставляют пустым. Однако, чтобы генерировать одинаковые последовательности случайных чисел, необходимо ввести число из диапазона от 1 до 32 767 (допускаются только целые числа). Тогда в будущем можно получить тот же набор выборочных значений, если в это поле снова ввести то же самое начальное значение.

Все диалоговые окна Генерация случайных чисел имеют область Параметры; опции этой области зависят от типа выбранного распределения. Назначение большинства этих опций очевидно, но некоторые требуют пояснений.

Равномерное распределение. Диалоговое окно Генерация случайных чисел для этого распределения показано на рис. 5.9.

Здесь в области Параметры надо задать только верхнюю и- нижнюю границы, в пределах которых сосредоточено распределение.

Нормальное распределение. Диалоговое окно Генерация случайных чисел для этого распределения показано на рис. 5.10.

В области Параметры задаются значения среднего (математического ожидания) и стандартное (среднеквадратическое отклонение). Для стандартного нормального распределения среднее равно 0, а стандартное отклонение — 1.

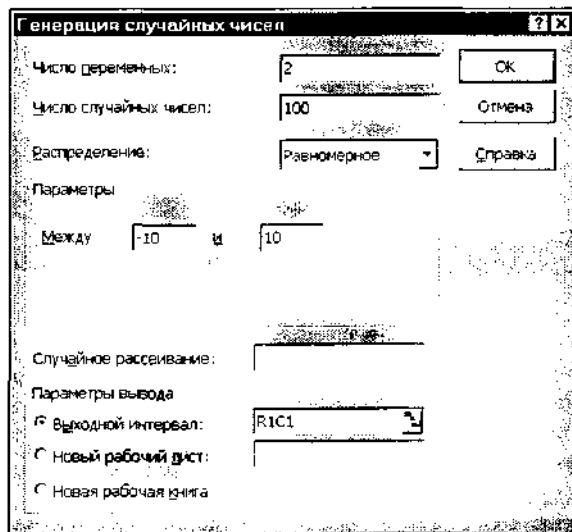


Рис. 5.9. Диалоговое окно для генерирования равномерно распределенных случайных чисел

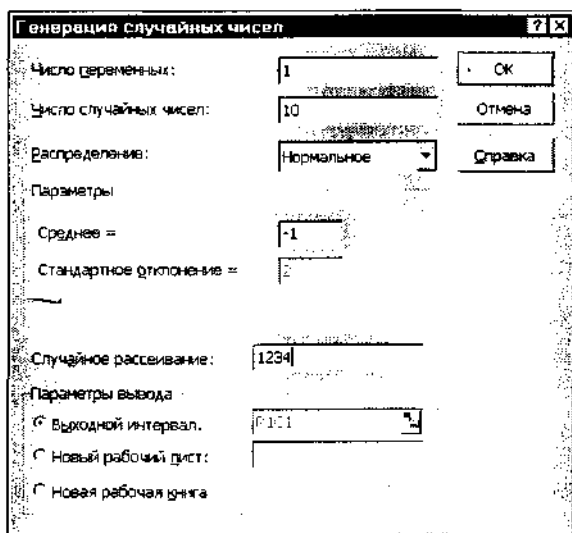


Рис. 5.10. Диалоговое окно для генерирования нормально распределенных случайных чисел

Распределение Бернулли. Диалоговое окно Генерация случайных чисел для данного случая показано на рис. 5.11.

Здесь в области Параметры задается только один параметр — вероятность p .

Биномиальное распределение. Диалоговое окно Генерация случайных чисел для этого распределения показано на рис. 5.12.

Для этого распределения задаются значения вероятности p и количество испытаний n .

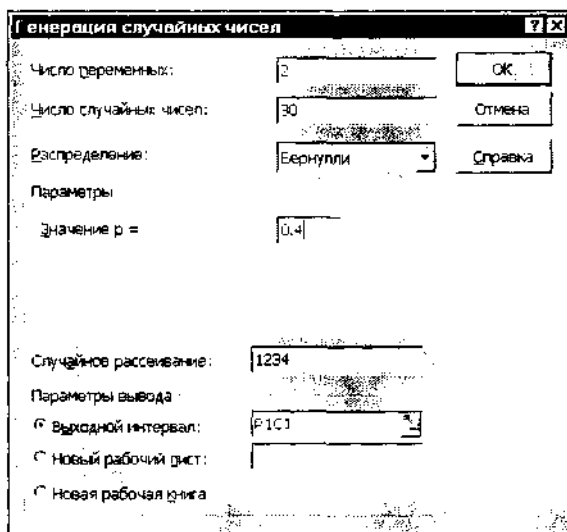


Рис. 5.11. Диалоговое окно для генерирования случайных чисел, имеющих распределение Бернулли

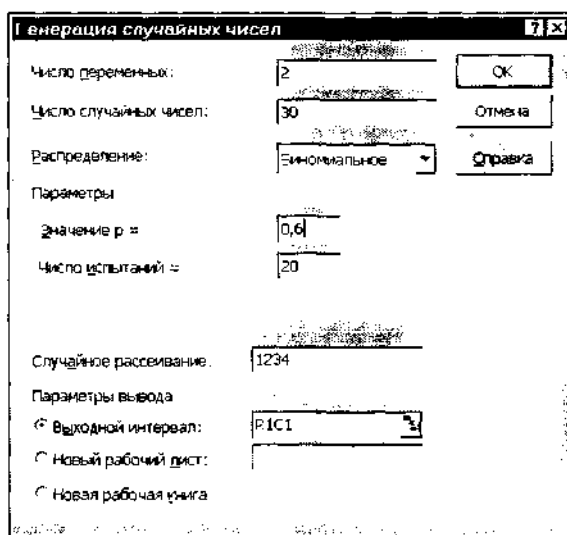


Рис. 5.12. Диалоговое окно для генерирования случайных чисел, имеющих биномиальное распределение

Распределение Пуассона. Диалоговое окно Генерация случайных чисел для данного случая показано на рис. 5.13.

Здесь в области Параметры задается только один параметр Лямбда.

Модельное распределение. Диалоговое окно Генерация случайных чисел для этого случая показано на рис. 5.14.

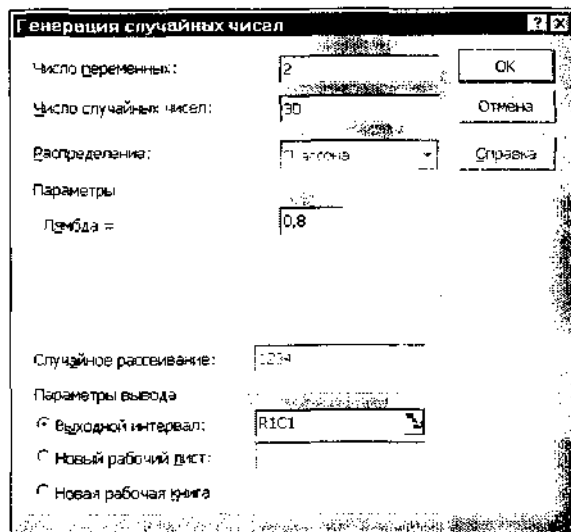


Рис. 5.13. Диалоговое окно для генерирования случайных чисел, имеющих распределение Пуассона

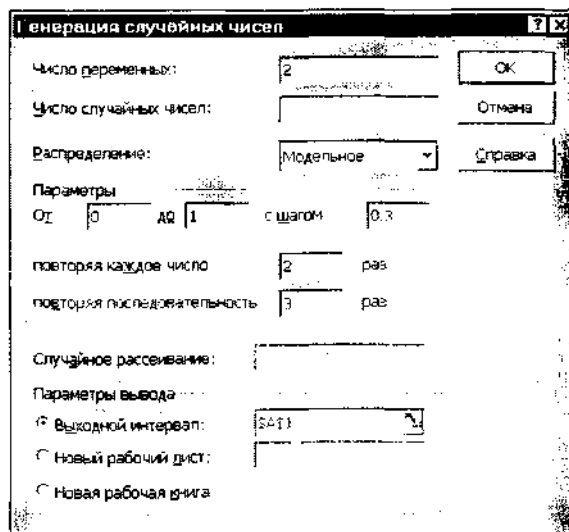


Рис. 5.14. Диалоговое окно для генерирования заданных чисел (модельное распределение)

Здесь задаются нижняя и верхняя границы чисел, шаг прогрессии, число повторений значений в последовательности и число повторений последовательно. На рис. 5.15 показаны сгенерированные числа с модельным распределением, параметры которого заданы на рис. 5.14.

	A	B	C
1	0	0	
2	0	0	
3	0,3	0,3	
4	0,3	0,3	
5	0,6	0,6	
6	0,6	0,6	
7	0,9	0,9	
8	0,9	0,9	
9	1	1	
10	1	1	
11	0	0	
12	0	0	
13	0,3	0,3	
14	0,3	0,3	
15	0,6	0,6	
16	0,6	0,6	
17	0,9	0,9	
18	0,9	0,9	
19	1	1	
20	1	1	
21	0	0	
22	0	0	

Рис. 5.15. Сгенерированные числа

Дискретное распределение. Диалоговое окно Генерация случайных чисел для этого типа распределения вместе с необходимыми входными данными показано на рис. 5.16.

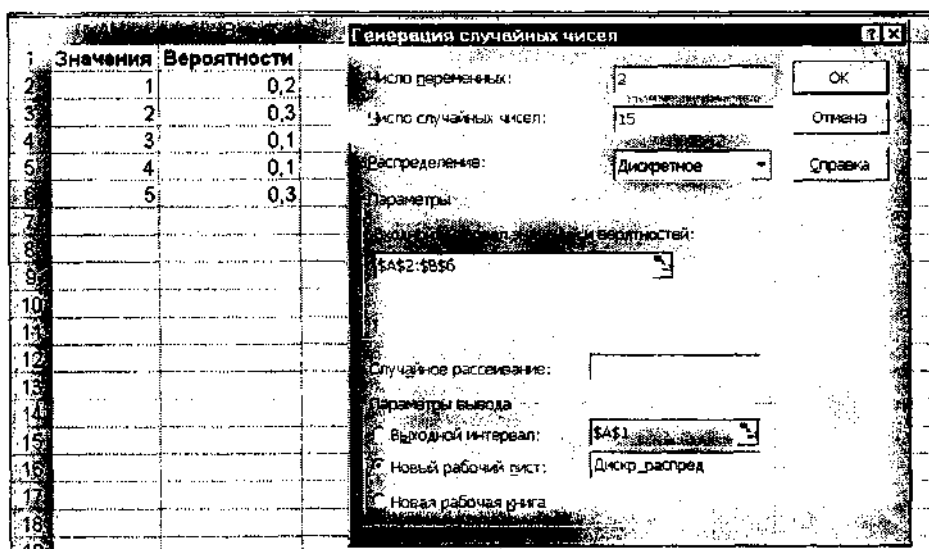


Рис. 5.16. Диалоговое окно для генерирования случайных чисел, имеющих заданное дискретное распределение

	A	B	C
1	4	1	
2	3	2	
3	5	2	
4	5	2	
5	5	1	
6	2	2	
7	4	2	
8	2	5	
9	2	5	
10	5	5	
11	5	1	
12	1	1	
13	4	5	
14	1	4	
15	2	2	
16			
17			

Рис. 5.17. Сгенерированные числа

Для задания дискретного распределения в поле Входной интервал значений и вероятностей необходимо указать адрес диапазона ячеек, содержащий значения случайной величины и соответствующие им вероятности. Диапазон должен состоять из двух столбцов: левого, содержащего значения, и правого, содержащего вероятности, как показано на рис. 5.16. Сумма вероятностей должна быть равна 1. На рис. 5.17 представлены сгенерированные числа с распределением, параметры которого заданы на рис. 5.16.

В заключение отметим, что в Excel имеются и другие средства генерирования случайных выборок, например функции СЛЧИС и СЛУЧМЕЖДУ (см. раздел 4.13). Подробно задача генерирования значений случайных величин рассмотрена в главе 7.

5.4. Выборка

Это средство из исходного числового множества выбирает указанное количество чисел, причем либо случайным образом, либо с заданным периодом (например, каждое второе или каждое десятое число). Такую операцию выбора числовых значений из заданного множества можно трактовать как создание выборки заданного объема, если исходное множество рассматривать как генеральную совокупность. Подобная операция часто составляет один из этапов предварительной обработки данных. Например, если исходная выборка слишком велика для обработки или построения диаграмм либо если исходные данные содержат периодическую составляющую, то можно создать выборку, содержащую значения только из отдельных частей периода.

5.4.1. Опции диалогового окна Выборка

Диалоговое окно Выборка показано на рис. 5.18. Адрес диапазона ячеек, содержащий исходный набор числовых значений, задается в поле Входной интервал. Если этот диапазон состоит из нескольких столбцов, то значения сначала будут извлекаться из первого столбца, затем из второго столбца и т.д. Средство Выборка откажется работать (выведет соответствующее окно предупреждения), если среди исходных данных имеются нечисловые значения.

В области Метод выборки необходимо указать, каким способом будут выбираться значения из исходного множества. Если установлен переключатель Периодический, то из исходного множества будет выбрано каждое n -е значение; число n задается в поле ввода Период. Количество выбранных значений будет равно количеству значений в исходном диапазоне, деленному на значение в поле Период. Если установлен переключатель Случайный, значения из исходного множества выбираются случайным образом; количество выбираемых значений задается в поле Число выборок.

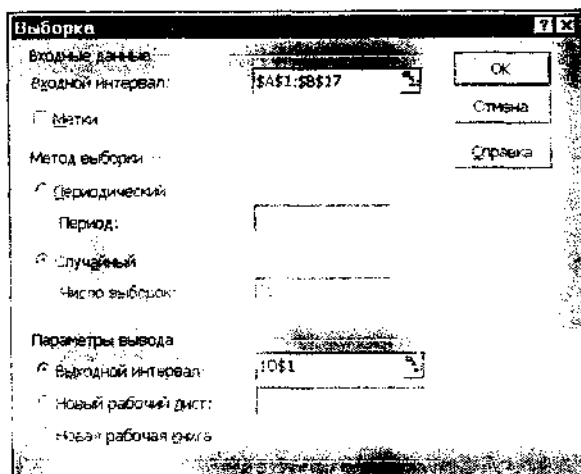


Рис. 5.18. Диалоговое окно Выборка

5.5. Ранг и перцентиль

Это средство позволяет создать таблицу, содержащую порядковый и процентный ранги для каждого значения в заданном наборе данных, при этом значения упорядочиваются в *порядке убывания*. На рис. 5.19 показаны диалоговое окно Ранг и перцентиль и исходные данные, на рис. 5.20 — результат применения этого средства. Итоговая таблица содержит порядковый номер выборочного значения, столбец выборочных значений, отсортированных в порядке убывания, столбец рангов и столбец процентных рангов этих значений, причем наибольшему значению присваивается ранг 1 и процентный ранг 100%, а наименьшему — наибольший ранг и процентный ранг, равный 0%.

Если имеется группа совпадающих значений, то им присваиваются одинаковые ранги, равные рангу первого числа из группы совпадающих значений. Значению, следующему за этой группой, присваивается ранг, больший ранга совпадающих значений на число этих одинаковых значений. Процентный ранг T_i для

выборочного значения x_i рассчитывается по формуле $T_i = \frac{n - R_i}{n - 1} \cdot 100\%$, где R_i —

ранг значения x_i , рассчитанный при условии упорядочивания данных по убыванию, n — объем выборки.

5.6. Двухвыборочный z-тест для средних

Это средство применяется для проверки гипотезы о равенстве (неравенстве) математических ожиданий двух независимых генеральных совокупностей, имеющих нормальное распределение, при известных дисперсиях этих распределений (см. раздел 2.4.2). Пусть имеются две независимые выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_m объемом соответственно n и m , извлеченные из совокупностей, имеющих нормальные распределения с известными дисперсиями σ_1^2 и σ_2^2 и неизвестными математическими ожиданиями соответственно μ_1 и μ_2 . Проверяется нулевая

	A	B	C	D	E	F	G	H
1	№ п/п	Выборка						
2	1	-0,300232						
3	2	1,2764735						
4	3	-2,183588						
5	4	-1,086701						
6	5	-1,846911						
7	6	-2,117931						
8	7	0,1348531						
9	8	-0,370241						
10	9	-0,186158						
11	10	0,865673						
12	11	1,6614558						
13	12	0,9021915						
14	13	-0,523795						
15	14	0,7576114						
16	15	-1,521571						
17	16	0,028117						
18	17	-1,742483						
19	18	1,44767						
20	19	0,7577137						
21	20	1,6614558						

Ранг и перцентиль

Входные данные: Вводный интервал: \$B\$1:\$B\$21

Группирование: ☐ по столбцам ☐ по строкам

☒ Метки в первой строке

Параметры вывода: ☒ Выводной интервал: \$D\$1 ☐ Новый рабочий лист ☐ Новая рабочая книга

OK Отмена Справка

Рис. 5.19. Исходные данные и диалоговое окно Ранг и перцентиль

	A	B	C	D	E	F
1	№ п/п	Выборка	Точка	Выборка	Ранг	Процент
2	1	-0,300232	11	1,661456	1	100,00%
3	2	1,2764735	18	1,44767	2	94,70%
4	3	-2,183588	2	1,276474	3	89,40%
5	4	-1,086701	12	0,902191	4	84,20%
6	5	-1,846911	10	0,865673	5	78,90%
7	6	-2,117931	19	0,757714	6	73,60%
8	7	0,1348531	14	0,757611	7	68,40%
9	8	-0,370241	20	0,25177	8	63,10%
10	9	-0,186158	7	0,134853	9	57,80%
11	10	0,865673	16	0,028117	10	52,60%
12	11	1,6614558	9	-0,18616	11	47,30%
13	12	0,9021915	1	-0,30023	12	42,10%
14	13	-0,523795	8	-0,37024	13	36,80%
15	14	0,7576114	13	-0,5238	14	31,50%
16	15	-1,521571	4	-1,0867	15	26,30%
17	16	0,028117	15	-1,52157	16	21,00%
18	17	-1,742483	17	-1,74248	17	15,70%
19	18	1,44767	5	-1,84691	18	10,50%
20	19	0,7577137	6	-2,11793	19	5,20%
21	20	0,25177	3	-2,18359	20	,00%

Рис. 5.20. Результат вычислений

гипотеза $H_0: \mu_1 - \mu_2 = \delta$ (δ задано). Z-тест позволяет проверить гипотезу H_0 против разных конкурирующих гипотез: $H_1: \mu_1 \neq \mu_2 + \delta$ или $H_1: \mu_1 > \mu_2 + \delta$, либо $H_1: \mu_1 < \mu_2 + \delta$. Критериальная статистика вычисляется по формуле

$$z = \frac{\bar{x} - \bar{y} - \delta}{\sqrt{\frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}}},$$

где \bar{x} и \bar{y} — выборочные средние соответственно первой и второй выборок.

Для выборок из нормально распределенных генеральных совокупностей критерияльная статистика z имеет стандартное нормальное распределение. Поэтому при заданном уровне значимости α критическая область строится на основе стандартного нормального распределения — вычисляется квантиль t порядка $1 - \alpha$ для проверки гипотезы о равенстве либо квантиль t порядка $1 - \alpha/2$ для проверки гипотез неравенства. Нулевая гипотеза о равенстве принимается, если $|z| < t$ (в противном случае отвергается); гипотеза H_0 при конкурирующей гипотезе $H_1: \mu_1 > \mu_2 + \delta$ принимается, если $z < t$; и при конкурирующей гипотезе $H_1: \mu_1 < \mu_2 + \delta$ нулевая гипотеза принимается при выполнении неравенства $-t < z$.

Рассмотрим пример. Имеется две выборки¹ объемом соответственно 50 и 20 значений, показанные на рис. 5.21. Обе имеют нормальное распределение, первая — стандартное (т.е. $\mu_1 = 0$ и $\sigma_1^2 = 1$), а для второй — $\mu_2 = 1$ и $\sigma_2^2 = 2$. Проверим с помощью средства Двухвыборочный z-тест для средних нулевую гипотезу, что $\mu_2 - \mu_1 = 1,5$ для разных случаев конкурирующих гипотез. Заполненное диалоговое окно для этого примера также показано на рис. 5.21.

	A	B	C	D	E	F	G	H	I
1	Выборка1	Выборка2		Двухвыборочный z-тест для средних					
2	-0,300232	5,9675145		Входные данные					
3	-1,277683	1,1974558		Интервал переменной 1: \$B\$1:\$B\$21					
4	0,2442573	-0,7640696		Интервал переменной 2: \$A\$1:\$A\$51					
5	1,2764735	-0,581916		Гипотетическая средняя разность: 1,5					
6	1,1983502	-0,6243212		Дисперсия переменной 1 (известная): 1					
7	1,7331331	1,1688795		Дисперсия переменной 2 (известная):					
8	-2,183588	1,4044341		<input checked="" type="checkbox"/> Метки					
9	-0,234181	-2,3911238		Альфа: 0,05					
10	1,0950225	0,0832612		Параметры вывода					
11	-1,086701	1,0600608		<input checked="" type="radio"/> Выходной интервал: 10%					
12	-0,690204	-0,3645285		<input type="radio"/> Новый рабочий лист:					
13	-1,690432	1,0935916		<input type="radio"/> Новая рабочая книга					
14	-1,846911	1,6365735							
15	-0,977629	5,5526031							
16	-0,773507	-2,829191							
17	-2,117931	0,9367265							
18	-0,567925	4,8187318							
19	0,404048	1,723457							

Рис. 5.21. Исходные данные и диалоговое окно Двухвыборочный z-тест для средних

Отметим, что средство требует, чтобы δ , значение которого задается в поле Гипотетическая средняя разность, было неотрицательно. Поэтому первым (в поле ввода Интервал переменной 1) задается адрес диапазона ячеек, содержащий выборку с большим математическим ожиданием, а затем в поле Интервал переменной 2

¹ Выборки получены с помощью средства Генерация случайных чисел.

указывается адрес второй выборки. В полях ввода Дисперсия переменной 1 и Дисперсия переменной 2 вводятся значения дисперсий соответственно первой и второй выборок. В поле Альфа вводится значение уровня значимости α . Результат вычислений средства Двухвыборочный z-тест для средних показан на рис. 5.22.

	A	B	C	D	E	F
1	Выборка1	Выборка2	Двухвыборочный z-тест для средних			
2	-0.300232	5.9675145				
3	-1.277683	1.1974558			Выборка2	Выборка1
4	0.2442573	-0.7640696	Среднее		0.992495873	-0.10700326
5	1.2764735	-0.581916	Известная дисперсия		2	1
6	1.1983502	-0.6243212	Наблюдения		20	50
7	1.7331331	1.1688795	Гипотетическая разность средних		1.5	
8	-2.183588	1.4044341	z		-1.15614643	
9	-0.234181	-2.3911238	P(Z<=z) одностороннее		0.123810689	
10	1.0950225	0.0832612	z критическое одностороннее		1.644853476	
11	-1.086701	1.0600608	P(Z<=z) двухстороннее		0.247621378	
12	-0.690204	-0.3645285	z критическое двухстороннее		1.959962787	
13	-1.690432	1.0935916				
14	-1.846911	1.6365735				
15	-0.977629	5.5526031				
16	-0.773507	-2.829191				
17	-2.117931	0.9367265				
18	-0.567925	4.8187318				

Рис. 5.22. Результат вычислений

В итоговой таблице приводятся следующие данные.

- Среднее — выборочные средние выборок.
- Известная дисперсия — дисперсии выборок, которые указаны в диалоговом окне.
- Наблюдения — объемы выборок.
- Гипотетическая разность средних — значение 5, которое задано в диалоговом окне.
- z — значение критериальной статистики.
- $P(Z \leq z)$ одностороннее — вероятность $P(X < z)$, где X — случайная величина, распределенная по стандартному нормальному закону, z — подсчитанное значение критериальной статистики.
- z критическое одностороннее — значение квантиля порядка $1 - \alpha/2$.
- $P(Z \leq z)$ двухстороннее — вероятность $P(|X| < |z|)$, где X — случайная величина, распределенная по стандартному нормальному закону, z — подсчитанное значение критериальной статистики.
- z критическое двухстороннее — значение квантиля порядка $1 - \alpha$.

Как видно из результатов расчета, в данном примере нет оснований отвергать нулевую гипотезу при любых конкурирующих гипотезах.

Статистическая функция ZТЕСТ (см. раздел 4.8.1) вычисляет вероятность $P(Z \leq z)$ двухстороннее.

5.7. Двухвыборочный t-тест с одинаковыми дисперсиями

Это средство реализует критерий проверки гипотезы о равенстве (неравенстве) математических ожиданий распределений двух независимых генеральных совокупностей, имеющих нормальные распределения с неизвестными дисперсиями в предположении, что дисперсии равны. Этот критерий, называемый t-тестом или тестом Стьюдента, подробно описан в разделе 2.4.2.

Рассмотрим выходные данные, вычисляемые этим средством, на примере проверки нулевой гипотезы $H_0: \mu_1 - \mu_2 = \delta$ (δ задано) против разных конкурирующих гипотез: $H_1: \mu_1 \neq \mu_2 + \delta$ или $H_1: \mu_1 > \mu_2 + \delta$, либо $H_1: \mu_1 < \mu_2 + \delta$ (μ_1 и μ_2 — неизвестные математические ожидания выборок). Исходные данные и заполненное диалоговое окно Двухвыборочный t-тест с одинаковыми дисперсиями показаны на рис. 5.23. Выборки извлечены из нормально распределенных генеральных совокупностей с одной и той же дисперсией, равной 1, и математическими ожиданиями 0 и 1 соответственно². Проверим гипотезу, что $\mu_2 - \mu_1 = 2$ (на самом деле $\mu_2 - \mu_1 = 1$).

	A	B	C	D	E	F	G	H	I
1	Выборка1	Выборка2							
2	-0,300232	0,6997678							
3	-1,277683	-0,277683							
4	0,2442573	1,2442573							
5	1,2764735	2,2764735							
6	1,1983502	2,1983502							
7	1,7331331	2,7331331							
8	-2,183588	-1,183588							
9	-0,234181	0,7658188							
10	1,0950225	2,0950225							
11	-1,086701	-0,086701							
12	-0,690204	0,3097958							
13	-1,690432	-0,690432							
14	-1,846911	-0,846911							
15	-0,977629	0,0223705							
16	-0,773507	0,2264929							
17	-2,117931	-1,117931							
18	-0,567925	0,4320751							

Двухвыборочный t-тест с одинаковыми дисперсиями

Входные данные

Интервал переменной 1:

Интервал переменной 2:

Гипотетическая средняя разность:

☒ Метки

Альфа:

Параметры вывода

☒ Выходной интервал:

☐ Новый рабочий лист:

☐ Новая рабочая книга

ОК Отмена Справка

Рис. 5.23. Исходные данные и диалоговое окно Двухвыборочный t-тест с одинаковыми дисперсиями

Отметим, что средство требует, чтобы 5, значение которого задается в поле Гипотетическая средняя разность, было неотрицательно. Поэтому первым (в поле ввода Интервал переменной 1) задается адрес диапазона ячеек, содержащий выборку с большим математическим ожиданием, а затем в поле Интервал переменной 2 указывается адрес второй выборки. (Диапазоны должны состоять из одного столбца или одной строки.) В поле Альфа вводится значение уровня значимости α . Результат вычислений средства Двухвыборочный t-тест с одинаковыми дисперсиями показан на рис. 5.24.

² Выборки получены с помощью средства Генерация случайных чисел.

	A	B	C	D	E	F
1	Выборка1	Выборка2		Двухвыборочный t-тест с одинаковыми дисперсиями		
2	-0,300232	0,6997678				
3	-1,277683	-0,277683			Выборка2	Выборка1
4	0,2442573	1,2442573		Среднее	0,8528339	-0,107003
5	1,2764735	2,2764735		Дисперсия	1,4100187	1,3535148
6	1,1983502	2,1983502		Наблюдения	30	50
7	1,7331331	2,7331331		Объединенная дисперсия	1,3745227	
8	-2,183588	-1,183588		Гипотетическая разность средних	2	
9	-0,234181	0,7658188		df	78	
10	1,0950225	2,0950225		t-статистика	-3,841723	
11	-1,086701	-0,086701		P(T<=t) одностороннее	0,0001237	
12	-0,690204	0,3097958		t критическое одностороннее	1,6646254	
13	-1,690432	-0,690432		P(T<=t) двухстороннее	0,0002474	
14	-1,846911	-0,846911		t критическое двухстороннее	1,9908475	
15	-0,977629	0,0223705				
16	-0,773507	0,2264929				
17	-2,117931	-1,117931				
18	-0,567925	0,4320751				

Рис. 5.24. Результат вычислений

В итоговой таблице приводятся следующие данные.

- Среднее — выборочные средние для каждой выборки.
- Дисперсия — несмещенные выборочные оценки дисперсий выборок.
- Наблюдения — объемы выборок.
- Гипотетическая разность средних — значение δ , которое задано в диалоговом окне.
- Объединенная дисперсия — “средняя” оценка дисперсии; рассчитывается по формуле $s^2 = \frac{(n-1)s_1^2 + (m-1)s_2^2}{n+m-2}$, где n и m — объемы выборок, s_i^2 — оценки дисперсий (их значения приводятся в строке Дисперсия).
- df — число степеней свободы; вычисляется как $n + m - 2$.
- t-статистика — значение критериальной статистики; вычисляется по формуле $t = \frac{\sqrt{n+m-2}(\bar{x} - \bar{y} - \delta)}{\sqrt{\frac{n+m}{nm} \sqrt{(n-1)s_1^2 + (m-1)s_2^2}}}$, имеет распределение Стьюдента с df степенями свободы.
- P(T<=t) одностороннее — вероятность $P(X \leq t)$, где X — случайная величина, имеющая распределение Стьюдента с df степенями свободы, t — подсчитанное значение критериальной статистики.
- t критическое одностороннее — значение квантиля $t_{\alpha/2}$ порядка $1 - \alpha$ распределения Стьюдента с df степенями свободы.

- $P(T \leq t)$ двухстороннее — вероятность $P(|X| \leq |t|)$, где X — случайная величина, имеющая распределение Стьюдента с df степенями свободы, t — подсчитанное значение критериальной статистики.
- t критическое двухстороннее — значение квантиля $t_{кр1}$ порядка $1 - \alpha/2$ распределения Стьюдента с df степенями свободы.

Нулевая гипотеза $H_0: \mu_1 - \mu_2 = \delta$ принимается, если $|t| < t_{кр1}$ (в противном случае отвергается); гипотеза H_0 при конкурирующей гипотезе $H_1: \mu_1 > \mu_2 + \delta$ принимается, если $t < t_{кр2}$; при конкурирующей гипотезе $H_1: \mu_1 < \mu_2 + \delta$ нулевая гипотеза принимается при выполнении неравенства $t_{кр2} < t$.

Как видно из результатов расчета, в данном примере нулевую гипотезу следует отвергнуть при любых конкурирующих гипотезах.

Статистическая функция ТТЕСТ при значении аргумента Тип = 2 (см. раздел 3.8.2) вычисляет вероятности $P(T \leq t)$ двухстороннее и $P(T \leq t)$ одностороннее.

5.8. Двухвыборочный t-тест с различными дисперсиями

Это средство реализует критерий проверки гипотезы о равенстве (неравенстве) математических ожиданий распределений двух независимых генеральных совокупностей, имеющих нормальные распределения с неизвестными и различными дисперсиями. Этот критерий также называется t -тестом или тестом Стьюдента для неравных дисперсий, либо критерием Фишера–Беренса и подробно описывается в разделе 2.4.2.

Рассмотрим выходные данные, вычисляемые этим средством, на примере проверки нулевой гипотезы $H_0: \mu_1 - \mu_2 = \delta$ (δ задано) против разных конкурирующих гипотез: $H_1: \mu_1 \neq \mu_2 + \delta$ или $H_1: \mu_1 > \mu_2 + \delta$, либо $H_1: \mu_1 < \mu_2 + \delta$ (μ_1 и μ_2 — неизвестные математические ожидания выборок). Повторим тест на примере данных из предыдущего раздела, т.е. выборки извлечены из нормально распределенных генеральных совокупностей с одной и той же дисперсией, равной 1, и математическими ожиданиями соответственно 0 и 1. Проверим гипотезу, что $\mu_2 - \mu_1 = 2$ (на самом деле $\mu_2 - \mu_1 = 1$). Исходные данные и заполненное диалоговое окно Двухвыборочный t-тест с различными дисперсиями показаны на рис. 5.25.

Отметим, что средство требует, чтобы δ , значение которого задается в поле Гипотетическая средняя разность, было неотрицательно. Поэтому первым (в поле ввода Интервал переменной 1) задается адрес диапазона ячеек, содержащий выборку с большим математическим ожиданием, а затем в поле Интервал переменной 2 указывается адрес второй выборки. (Диапазоны должны состоять из одного столбца или одной строки.) В поле Альфа вводится значение уровня значимости α . Результат вычислений средства Двухвыборочный t-тест с различными дисперсиями показан на рис. 5.26.

В итоговой таблице приводятся следующие данные.

- Среднее — выборочные средние для каждой выборки.
- Дисперсия — несмещенные выборочные оценки дисперсий выборок.

	A	B	C	D	E	F	G	H
1	Выборка1	Выборка2						
2	-0,300232	0,6997678						
3	-1,277683	-0,277683						
4	0,2442573	1,2442573						
5	1,2764735	2,2764735						
6	1,1983502	2,1983502						
7	1,7331331	2,7331331						
8	-2,183588	-1,183588						
9	-0,234181	0,7658188						
10	1,0950225	2,0950225						
11	-1,086701	-0,086701						
12	-0,690204	0,3097958						
13	-1,690432	-0,690432						
14	-1,846911	-0,846911						
15	-0,977629	0,0223705						
16	-0,773507	0,2264929						
17	-2,117931	-1,117931						
18	-0,567925	0,4320751						
19	0,404048	0,5959512						

Двухвыборочный t-тест с различными дисперсиями

Входные данные

Интервал переменной 1: \$B\$1:\$B\$31

Интервал переменной 2: \$A\$1:\$A\$51

Гипотетическая средняя разность: 8

Р-уровень: 0,05

Параметры вывода

☒ Выходной интервал: \$D\$1

☐ Новый рабочий лист

☐ Новая рабочая книга

OK Отмена Справка

Рис. 5.25. Исходные данные и диалоговое окно Двухвыборочный t-тест с различными дисперсиями

	A	B	C	D	E	F
1	Выборка1	Выборка2		Двухвыборочный t-тест с различными дисперсиями		
2	-0,300232	0,6997678				
3	-1,277683	-0,277683			Выборка2	Выборка1
4	0,2442573	1,2442573		Среднее	0,852833942	-0,107003257
5	1,2764735	2,2764735		Дисперсия	1,410018743	1,353514772
6	1,1983502	2,1983502		Наблюдения	30	50
7	1,7331331	2,7331331		Гипотетическая разность средних	2	
8	-2,183588	-1,183588		df	60	
9	-0,234181	0,7658188		t-статистика	-3,821883531	
10	1,0950225	2,0950225		P(T<=t) одностороннее	0,000158539	
11	-1,086701	-0,086701		t критическое одностороннее	1,670648544	
12	-0,690204	0,3097958		P(T<=t) двухстороннее	0,000317078	
13	-1,690432	-0,690432		t критическое двухстороннее	2,000297172	
14	-1,846911	-0,846911				
15	-0,977629	0,0223705				
16	-0,773507	0,2264929				
17	-2,117931	-1,117931				
18	-0,567925	0,4320751				

Рис. 5.26. Результат вычислений

- Наблюдения — объемы выборок.
- Гипотетическая разность средних — значение 8, которое задано в диалоговом окне.

- df — число степеней свободы; вычисляется по формуле $\frac{\left(\frac{s_1^2}{n} + \frac{s_2^2}{m}\right)^2}{\frac{(s_1^2/n)^2}{n-1} + \frac{(s_2^2/m)^2}{m-1}}$, где

s_1^2 и s_2^2 — несмещенные оценки дисперсий (их значения приводятся в строке Дисперсия), n и m — объемы соответственно первой и второй выборок.

- t -статистика — значение критериальной статистики; вычисляется по формуле $t = \frac{\bar{x} - \bar{y} - \delta}{\sqrt{\frac{s_1^2}{n} + \frac{s_2^2}{m}}}$, имеет распределение, близкое к распределению Стюдента с df степенями свободы.

- $P(T \leq t)$ одностороннее — вероятность $P(X \leq t)$, где X — случайная величина, имеющая распределение Стюдента с df степенями свободы, t — подсчитанное значение критериальной статистики.
- t критическое одностороннее — значение квантиля $t_{кр2}$ порядка $1 - \alpha$ распределения Стюдента с df степенями свободы.
- $P(T \leq t)$ двухстороннее — вероятность $P(|X| \leq |t|)$, где X — случайная величина, имеющая распределение Стюдента с df степенями свободы, t — подсчитанное значение критериальной статистики.
- t критическое двухстороннее — значение квантиля $t_{кр1}$ порядка $1 - \alpha/2$ распределения Стюдента с df степенями свободы.

Нулевая гипотеза $H_0: \mu_1 - \mu_2 = \delta$ принимается, если $|t| < t_{кр1}$ (в противном случае отвергается); гипотеза H_0 при конкурирующей гипотезе $H_1: \mu_1 > \mu_2 + \delta$ принимается, если $t < t_{кр2}$; при конкурирующей гипотезе $H_1: \mu_1 < \mu_2 + \delta$ нулевая гипотеза принимается при выполнении неравенства $t_{кр2} < t$.

Как видно из результатов расчета, в данном примере нулевую гипотезу следует отвергнуть при любых конкурирующих гипотезах.

Статистическая функция ТТЕСТ при значении аргумента Тип = 3 (см. раздел 3.8.2) вычисляет вероятности $P(T \leq t)$ двухстороннее и $P(T \leq t)$ одностороннее.

5.9. Парный двухвыборочный t -тест для средних

Это средство реализует критерий проверки гипотезы о равенстве (неравенстве) математических ожиданий распределений двух зависимых выборок, имеющих нормальные распределения. Этот критерий также называется t -тестом или тестом Стюдента для парных наблюдений и подробно описан в разделе 2.4.2.

Рассмотрим выходные данные, вычисляемые этим средством, на примере проверки нулевой гипотезы $H_0: \mu_1 - \mu_2 = \delta$ (δ задано) против разных конкурирующих гипотез: $H_1: \mu_1 \neq \mu_2 + \delta$ или $H_1: \mu_1 > \mu_2 + \delta$, либо $H_1: \mu_1 < \mu_2 + \delta$ (μ_1 и μ_2 — неизвестные математические ожидания выборок). Рассмотрим пример, когда выборки извлечены из нормально распределенных генеральных совокупностей с математическими ожиданиями соответственно 0 и 1.

Проверим гипотезу, что $|j_2 - j_1| = 1,5$ (на самом деле $\mu_2 - \mu_1 = 1$). Исходные данные и заполненное диалоговое окно Парный двухвыборочный t-тест для средних показаны на рис. 5.27.

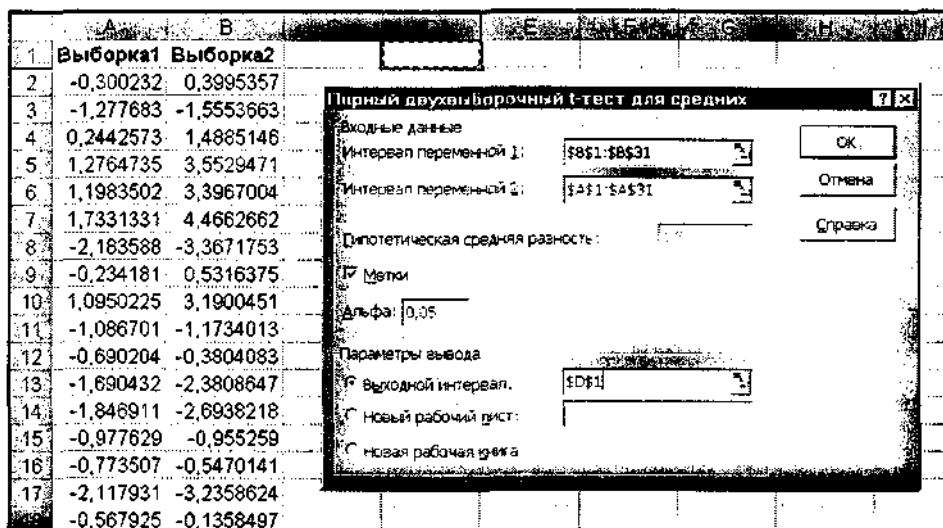


Рис. 5.27. Исходные данные и диалоговое окно Парный двухвыборочный t-тест для средних

Отметим, что средство требует, чтобы 5, значение которого задается в поле Гипотетическая средняя разность, было неотрицательно. Поэтому первым (в поле ввода Интервал переменной 1) задается адрес диапазона ячеек, содержащий выборку с большим математическим ожиданием, а затем в поле Интервал переменной 2 указывается адрес второй выборки. (Диапазоны должны состоять из одного столбца или одной строки.) В поле Альфа вводится значение уровня значимости α . Результат вычислений средства Парный двухвыборочный t-тест для средних показан на рис. 5.28.

В итоговой таблице приводятся следующие данные.

- Среднее — выборочные средние для каждой выборки.
- Дисперсия — несмещенные выборочные оценки дисперсий выборок.
- Наблюдения — объемы выборок.
- Корреляция Пирсона — выборочный коэффициент корреляции; вычисляется

по формуле
$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}$$

Гипотетическая разность средних значение 5, которое задано в Диалоговом окне.

	A	B	C	D	E	F
1	Выборка1	Выборка2	Парный двухвыборочный t-тест для средних			
2	-0,300232	0,3995357				
3	-1,277683	-1,5553663		Выборка2	Выборка1	
4	0,2442573	1,4885146	Среднее	0,937821973	-0,147166058	
5	1,2764735	3,5529471	Дисперсия	9,684347868	1,410018743	
6	1,1983502	3,3967004	Наблюдения	30	30	
7	1,7331331	4,4662662	Корреляция Пирсона	0,489593926		
8	-2,183588	-3,3671753	Гипотетическая разность средних	1,5		
9	-0,234181	0,5316375	df	29		
10	1,0950225	3,1900451	t-статистика	-0,831355674		
11	-1,086701	-1,1734013	P(T<=t) одностороннее	0,206282928		
12	-0,690204	-0,3804083	t критическое одностороннее	1,699127097		
13	-1,690432	-2,3808647	P(T<=t) двухстороннее	0,412565857		
14	-1,846911	-2,6938218	t критическое двухстороннее	2,045230758		
15	-0,977629	-0,955259				
16	-0,773507	-0,5470141				
	-2,117931	-3,2358624				

Рис. 5.28. Результат вычислений

- df — число степеней свободы, равное $n - 1$.
- t -статистика — значение критериальной статистики; вычисляется по формуле $t = \frac{\bar{d} - \delta}{S_n / \sqrt{n}}$, где $\bar{d} = \frac{1}{n} \sum_{i=1}^n (x_i - y_i)$, $S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - y_i - \bar{d})^2$, и имеет распределение Стьюдента с df степенями свободы.
- $P(T \leq t)$ одностороннее — вероятность $P(X \leq t)$, где X — случайная величина, имеющая распределение Стьюдента с df степенями свободы, t — подсчитанное значение критериальной статистики.
- t критическое одностороннее — значение квантиля $t_{кр2}$ порядка $1 - \alpha$ распределения Стьюдента с df степенями свободы.
- $P(T \leq t)$ двухстороннее — вероятность $P(|X| \leq t)$, где X — случайная величина, имеющая распределение Стьюдента с df степенями свободы, t — подсчитанное значение критериальной статистики.
- t критическое двухстороннее — значение квантиля $t_{кр1}$ порядка $1 - \alpha/2$ распределения Стьюдента с df степенями свободы.

Нулевая гипотеза $H_0: \mu_1 - \mu_2 = \delta$ принимается, если $|t| < t_{кр1}$ (в противном случае отвергается); гипотеза H_0 при конкурирующей гипотезе $H_1: \mu_1 > \mu_2 + \delta$ принимается, если $t < t_{кр2}$; при конкурирующей гипотезе $H_1: \mu_1 < \mu_2 + \delta$ нулевая гипотеза принимается при выполнении неравенства $t_{кр2} < t$.

Как видно из результатов расчета, в данном примере нулевую гипотезу следует принять при любых конкурирующих гипотезах.

Статистическая функция ТТЕСТ при значении аргумента Тип = 1 (см. раздел 3.8.2) вычисляет вероятности $P(T \leq t)$ двухстороннее и $P(T \leq t)$ одностороннее.

5.10. Двухвыборочный F-тест для дисперсий

Это средство реализует критерий Фишера проверки равенства дисперсий двух независимых выборок из нормально распределенных генеральных совокупностей с дисперсиями соответственно σ_1^2 и σ_2^2 . Критерий подробно описан в разделе 2.4.2.

Рассмотрим выходные данные, вычисляемые этим средством, на примере проверки нулевой гипотезы $H_0: \sigma_1^2 = \sigma_2^2$ против конкурирующей гипотезы $H_1: \sigma_1^2 \neq \sigma_2^2$. Рассмотрим пример, когда выборки извлечены из нормально распределенных генеральных совокупностей с равными дисперсиями 1,5. Исходные данные и заполненное диалоговое окно Двухвыборочный F-тест для дисперсий показаны на рис. 5.29.

	A	B	C	D	E	F	G	H
1	Выборка1	Выборка2						
2	-0,300232	0,6997678						
3	-1,277683	-0,277683						
4	0,2442573	1,2442573						
5	1,2764735	2,2764735						
6	1,1983502	2,1983502						
7	1,7331331	2,7331331						
8	-2,183588	-1,183588						
9	-0,234181	0,7658188						
10	1,0950225	2,0950225						
11	-1,086701	-0,086701						
12	-0,690204	0,3097958						
13	-1,690432	-0,690432						
14	-1,846911	-0,846911						
15	-0,977629	0,0223705						
16	-0,773507	0,2264929						
17	-2,117931	-1,117931						
18	-0,567925	0,4320751						

Рис. 5.29. Исходные данные и диалоговое окно Двухвыборочный F-тест для дисперсий

Отметим, что первой (в поле Входной интервал 1) должна задаваться выборка, имеющая большую дисперсию. В поле Альфа вводится значение уровня значимости α . Результат вычислений средства Двухвыборочный F-тест для дисперсий показан на рис. 5.30.

В итоговой таблице приводятся следующие данные.

- Среднее — выборочные средние для каждой выборки.
- Дисперсия — несмещенные выборочные оценки дисперсий выборок.
- Наблюдения — объемы выборок.
- df — числа степеней свободы, равные $n - 1$ и $m - 1$; n и m — объемы выборок.
- F — значение критериальной статистики, вычисляемой по формуле

$$F = \frac{S_x^2}{S_y^2}, \quad \text{где} \quad S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad S_y^2 = \frac{1}{m} \sum_{j=1}^m (y_j - \bar{y})^2, \quad \text{и имеющей } F\text{-}$$

распределение со степенями свободы $k_1 = n - 1$ и $k_2 = m - 1$ (о F-распределении речь идет в разделе 1.5.7).

	A	B	C	D	E	F
1	Выборка1	Выборка2	Двухвыборочный F-тест для дисперсии			
2	-0,300232	0,6997678				
3	-1,277683	-0,277683			Выборка2	Выборка1
4	0,2442573	1,2442573	Среднее		0,852833942	-0,107003257
5	1,2764735	2,2764735	Дисперсия		1,410018743	1,353514772
6	1,1983502	2,1983502	Наблюдения		30	50
7	1,7331331	2,7331331	df		29	49
8	-2,183588	-1,183588	F		1,041746106	
9	-0,234181	0,7658188	P(F<=f) одностороннее		0,439884678	
10	1,0950225	2,0950225	F критическое одностороннее		1,698879259	
11	-1,086701	-0,086701				
12	-0,690204	0,3097958				
13	-1,690432	-0,690432				
14	-1,846911	-0,846911				
15	-0,977629	0,0223705				
16	-0,773507	0,2264929				

Рис. 5.30. Результат вычислений

- $P(F \leq f)$ одностороннее — вероятность $P(X \leq F)$, где X — случайная величина, имеющая F -распределение с df степенями свободы, F — подсчитанное значение критериальной статистики.
- F критическое одностороннее — значение квантиля t порядка $1 - \alpha$ F -распределения с df степенями свободы.

Нулевая гипотеза H_0 принимается, если $F < t$ (в противном случае отвергается). Как видно из результатов расчета, в данном примере нулевую гипотезу следует принять.

Статистическая функция ФТЕСТ (см. раздел 4.8.3) вычисляет удвоенную вероятность $P(F \leq f)$ одностороннее.

5.11. Однофакторный дисперсионный анализ

Это средство реализует критерий проверки гипотезы о равенстве математических ожиданий нескольких независимых выборок, построенный на основе дисперсионного анализа. Однофакторный дисперсионный анализ описан в разделе 3.4.2. Здесь покажем применение средства Однофакторный дисперсионный анализ и опишем его выходные данные.

На рис. 5.31 показаны три выборки, имеющие нормальное распределение с математическими ожиданиями 0, 0,5 и 1 и среднеквадратическими отклонениями 1, 2 и 3 соответственно. Объемы выборок — 50, 40 и 30 значений. (Выборки сгенерированы с помощью средства Генерация случайных чисел.) На рис. 5.31 также показано заполненное диалоговое окно Однофакторный дисперсионный анализ. Обращаем внимание, что все три выборки задаются в виде одного диапазона ячеек. В случае, когда выборки имеют разные размеры, диапазон задается в соответствии с наибольшей выборкой и неизбежно содержит пустые ячейки. Но средство правильно определяет объемы выборок. Также отметим, что в данном случае результаты анализа будут выводиться на отдельный рабочий лист с именем Результаты, который автоматически вставится в текущую рабочую книгу.

	A	B	C	D	E	F	G	H	I
1	Выборка1	Выборка2	Выборка3						
2	-0,300232	5,4675145	2,8387391	Однофакторный дисперсионный анализ					
3	-1,277683	0,6974558	-1,77098	Входные данные					
4	0,2442573	-1,26407	2,7289108	Входной интервал: \$A\$1:\$C\$51					
5	1,2764735	-1,081916	3,839206	Группирование: <input type="radio"/> по столбцам					
6	1,1983502	-1,124321	2,2869782	<input type="radio"/> по строкам					
7	1,7331331	0,6688795	-3,444278	<input checked="" type="checkbox"/> Метки в первой строке					
8	-2,183588	0,9044341	-0,25585	Альфа: <input type="text"/>					
9	-0,234181	-2,891124	4,9809402	Параметры вывода					
10	1,0950225	-0,416739	-2,223076	<input type="checkbox"/> Выходной интервал: <input type="text"/>					
11	-1,086701	0,5600608	4,3497205	<input type="checkbox"/> Новый рабочий лист:					
12	-0,690204	-0,864529	4,4165328	<input type="checkbox"/> Новая рабочая книга:					
13	-1,690432	0,5935916	3,7801343						
14	-1,846911	1,1365735	-4,843021						
15	-0,977629	5,0526031	3,4056499						
16	-0,773507	-3,329191	-0,679293						
17	-2,117931	0,4367265	-1,309082						
18	-0,567925	4,3187318	-1,046506						

Рис. 5.31. Исходные данные и диалоговое окно Однофакторный дисперсионный анализ

На рис. 5.32 показаны результаты, выводимые средством Однофакторный дисперсионный анализ. Они представлены в виде двух таблиц, озаглавленных ИТОГИ и Дисперсионный анализ. В таблице ИТОГИ выводятся основные статистические характеристики выборок: в столбце Счет — объемы выборок, в столбце Сумма — суммы выборочных значений, в столбцах Среднее и Дисперсия — соответственно выборочные средние и дисперсии.

	A	B	C	D	E	F	G
1	Однофакторный дисперсионный анализ						
2							
3	ИТОГИ						
4	Группы	Счет	Сумма	Среднее	Дисперсия		
5	Выборка1	50	-5,3502	-0,107	1,35351477		
6	Выборка2	40	33,1763	0,829408	4,11373951		
7	Выборка3	30	36,8944	1,229813	7,23515686		
8							
9							
10	Дисперсионный анализ						
11	Источник вариации	SS	df	MS	F	P-Значение	F критическое
12	Между группами	38,5561	2	19,27807	5,16640019	0,007077273	3,073765242
13	Внутри групп	436,578	117	3,731433			
14							
15	Итого	475,134	119				
16							
17							

Рис. 5.32. Результат вычислений

Значения в первых четырех столбцах таблицы Дисперсионный анализ повторяют значения из дисперсионной таблицы (см. раздел 3.4.2). В столбце SS приведены суммы квадратов (межгрупповая, внутригрупповая и полная); в столбце df —

значения степеней свободы, а в столбце *MS* — дисперсии, межгрупповая и внутригрупповая. В столбце *F* записано значение критериальной статистики, в столбце *P-Значение* — значение вероятности $P(X > x)$, где *X* — случайная величина, имеющая χ^2 -распределение с *df* степенями свободы (о χ^2 -распределении речь идет в разделе 1.5.7). В столбце *Fкритическое* приводится критическое значение *t*, рассчитанное в соответствии с заданным уровнем значимости (параметр Альфа). Формулы для вычисления всех перечисленных значений приведены в разделе 3.4.2.

Нулевая гипотеза о равенстве математических ожиданий всех выборок принимается, если выполняется неравенство $F < F_{\text{критическое}}$. В нашем примере эту гипотезу следует отвергнуть.

5.12. Двухфакторный дисперсионный анализ с повторениями

Двухфакторный дисперсионный анализ описан в разделе 3.5.3. Здесь рассмотрим структуру входных данных для работы с этим средством и опишем выходные результаты. Структура входных данных представлена на рис. 5.33 (обозначения и пояснения приведены в разделе 3.5.3): в строке 1 показаны обозначения уровней фактора *p*; в столбце *A* — обозначения уровней фактора *u*; в данном случае имеется три выборки, поэтому под общим обозначением уровней фактора *u* записаны три строки числовых данных. Таким образом, в диапазоне, например, C8:C10 содержатся выборочные значения, соответствующие второму уровню фактора *P* и третьему уровню фактора *u*.

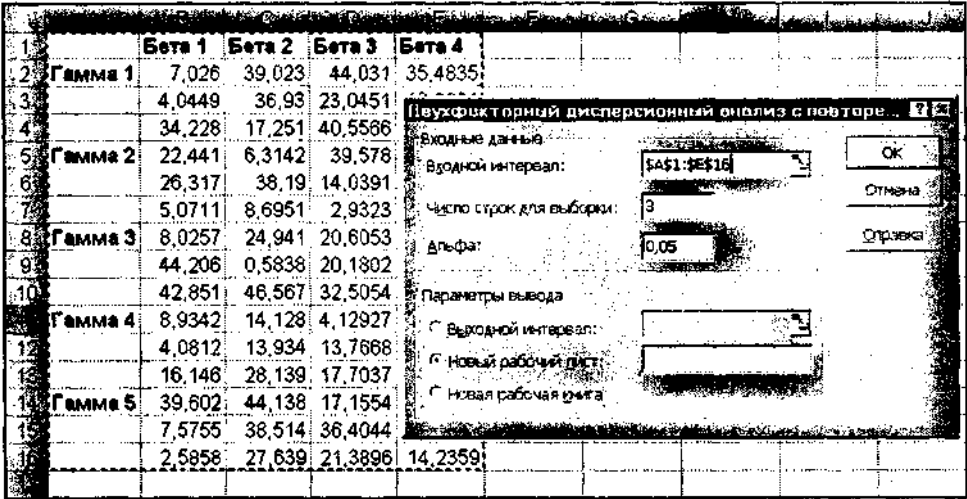


Рис. 5.33. Исходные данные и диалоговое окно Двухфакторный дисперсионный анализ с повторениями

Диалоговое окно рассматриваемого здесь средства показано на рис. 5.33. В поле Входной интервал указывается диапазон ячеек, содержащий входные данные, включая заголовки. В поле Число строк для выборки указывается ко-

личество рассматриваемых выборок, в данном случае введено число 3. В поле Альфа, как обычно, указывается значение уровня значимости.

На рис. 5.34 показаны выходные результаты работы данного средства, выведенные на отдельный рабочий лист. Выходные результаты сгруппированы в несколько таблиц. В первой таблице, озаглавленной ИТОГИ и состоящей из нескольких подтаблиц (по количеству уровней фактора u), приводятся статистические характеристики выборочных значений, соответствующих каждому сочетанию уровней фактора B и фактора u : количество выборочных значений (строка Счет), сумма выборочных значений (строка Сумма), выборочное среднее (строка Среднее) и выборочная дисперсия (строка Дисперсия). На рис. 5.34 показана такая подтаблица для первого уровня фактора u (таблица обозначена как Гамма 1), другие подобные подтаблицы, соответствующие другим уровням фактора u , на этом рисунке не показаны. В столбце Итого подтаблиц выводятся, такие же статистические характеристики выборочных значений, соответствующие одному уровню фактора u : количество выборочных значений, выборочное среднее и выборочная дисперсия (вычисляется по всем значениям данного уровня относительно общего среднего). В конце таблицы ИТОГИ выводится подтаблица Итого, в которой приведены те же характеристики, но подсчитанные по выборочным значениям для каждого уровня фактора B .

	A	B	C	D	E	F	G
1	Двухфакторный дисперсионный анализ с повторениями						
3	ИТОГИ	Бета 1	Бета 2	Бета 3	Бета 4	Итого	
4	Гамма 1						
5	Счет	3	3	3	3	12	
6	Сумма	45,299209	93,204725	107,63278	54,691573	300,8282832	
7	Среднее	15,099736	31,068242	35,877592	18,230524	25,0690236	
8	Дисперсия	276,64868	144,27867	126,5216	275,94289	231,244209	
34	Итого						
35	Счет	15	15	15	15		
36	Сумма	273,13536	384,98774	346,02225	322,19117		
37	Среднее	18,209024	25,66585	23,201483	21,479412		
38	Дисперсия	237,85875	217,7876	164,11801	244,1952		
41	Дисперсионный анализ						
42	Источник вариации	SS	df	MS	F	P-Значение	F критическое
43	Выборка	416,05598	4	104,014	0,5087581	0,726830298	2,605972327
44	Столбцы	441,70963	3	147,23654	0,7215858	0,545028464	2,838746127
45	Взаимодействие	3517,5466	12	293,12888	1,4365839	0,190046828	2,003480509
46	Внутри	8161,8313	40	204,04578			
47							
48	Итого	12537,143	59				

Рис. 5.34. Выходные результаты работы средства Двухфакторный дисперсионный анализ с повторениями

В нижней части выходных результатов приведена дисперсионная таблица (обозначения и вычисляющие формулы даны в разделе 3.5.3). Здесь в первом столбце, обозначенном SS, выведены суммы квадратов: соответственно SS_1 , SS_2 , SS_3 , SS_4 и в строке Итого — SS . В столбце df приведены степени свободы сумм квадратов, а в столбце MS — значения соответствующих дисперсий. В столбце F вычислены значения критериальных статистик, т.е. отношения дисперсий s_1^2 , s_2^2 , s_3^2 к дисперсии s_4^2 .

В столбце Р-Значение вычисляются вероятности $P(X > F)$, где X — случайная величина, имеющая F-распределение со степенями свободы, значения которых приведены в столбце df: первое значение степени свободы — из соответствующей строки этого столбца, а второе — всегда из четвертой строки, F — значение из столбца F. Например, значение в ячейке F43 (см. рис. 5.34), можно вычислить по формуле Excel `=FРАСП(E43;C43;C46)`. Эти значения используются для проверки гипотез о значимом влиянии факторов или их взаимного влияния: если вероятность больше заданного уровня значимости, то нулевая гипотеза об отсутствии влияния принимается, в противном случае — отвергается.

В столбце F критическое вычисляются критические значения, соответствующие заданному в диалоговом окне Двухфакторный дисперсионный анализ с повторениями уровню значимости α . Эти значения вычисляются как квантили порядка $1 - \alpha$ - F -распределения со степенями свободы, значения которых определяются так же, как при вычислении вероятностей из столбца Р-Значение. Например, значение в ячейке G43 (см. рис. 5.34) можно вычислить по формуле Excel `=FРАСПОБР(0,05;C43;C46)`. Эти значения используются для проверки гипотез о значимом влиянии факторов или их взаимного влияния: если значение в этом столбце больше значения в столбце F той же строки, то нулевая гипотеза об отсутствии влияния принимается, в противном случае — отвергается. Здесь принимаются все три нулевые гипотезы об отсутствии влияния факторов Р и у и их взаимного влияния. Однако значение в столбце F третьей строки (соответствует взаимному влиянию факторов) значительно больше аналогичных значений для отдельных факторов, и на это необходимо обратить внимание.

5.13. Двухфакторный дисперсионный анализ без повторений

Двухфакторный дисперсионный анализ описан в разделе 3.5.3. Структура входных данных показана на рис. 5.35 (обозначения и пояснения даны в разделе 3.5.3): в строке 1 приводятся обозначения уровней фактора Р; в столбце А — обозначения уровней фактора у; в диапазоне, обозначенном этими заголовками, введены числовые данные.

Диалоговое окно этого средства показано на рис. 5.35. В поле Входной интервал указывается диапазон ячеек, содержащий входные данные; если в этот диапазон включены заголовки строк и столбцов, то следует установить флажок опции Метки. В поле Альфа указывается значение уровня значимости.

На рис. 5.36 представлены выходные результаты работы данного средства, выведенные на отдельный рабочий лист. Выходные результаты сгруппированы в две таблицы. В первой таблице, озаглавленной ИТОГИ, приводятся статистические характеристики выборочных значений, соответствующих каждому уровню фактора Р (группировка по столбцам) и каждому уровню фактора у (группировка по строкам): количество выборочных значений (столбец Счет), сумма выборочных значений (столбец Сумма), выборочное среднее (столбец Среднее) и выборочная дисперсия (столбец Дисперсия).

В нижней части выходных результатов приведена дисперсионная таблица (обозначения и вычисляющие формулы даны в разделе 3.5.3). Здесь в первом

столбце, обозначенном SS, выведены суммы квадратов: соответственно SS_1 , SS_2 , SS_3 и в строке Итого — SS . В столбце df приведены степени свободы сумм квадратов, а в столбце MS — значения соответствующих дисперсий. В столбце F вычислены значения критериальных статистик, т.е. отношения дисперсий s_1^2 и s_2^2 к дисперсии s_j^2 .

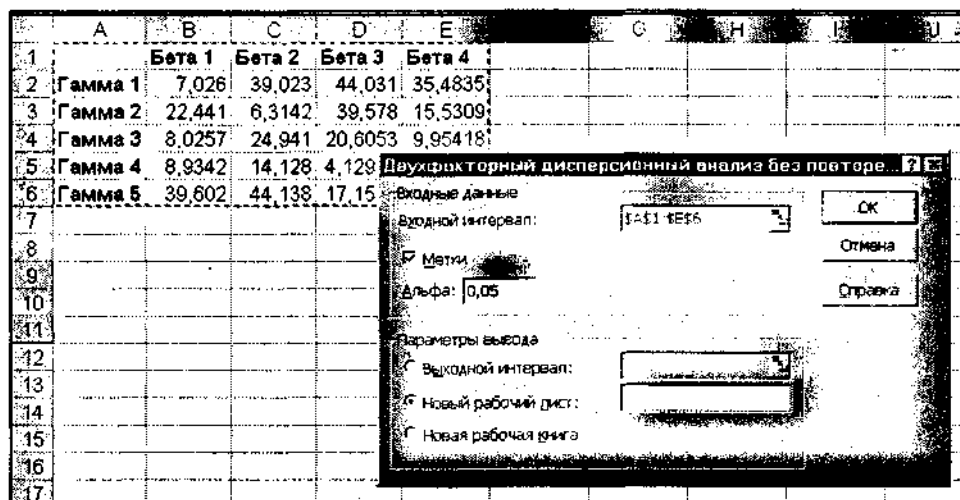


Рис. 5.35. Исходные данные и диалоговое окно Двухфакторный дисперсионный анализ без повторений

	A	B	C	D	E	F	G
1	Двухфакторный дисперсионный анализ без повторений						
3	ИТОГИ	Счет	Сумма	Среднее	Дисперсия		
4	Гамма 1	4	125,5637	31,39091	278,140477		
5	Гамма 2	4	83,86401	20,966	187,599247		
6	Гамма 3	4	63,52634	15,88158	87,0796283		
7	Гамма 4	4	69,89708	17,22427	283,047608		
8	Гамма 5	4	110,198	27,54949	287,150456		
10	Бета 1	5	86,02814	17,20563	196,333937		
11	Бета 2	5	128,5443	25,70885	256,94181		
12	Бета 3	5	125,499	25,0998	272,771505		
13	Бета 4	5	111,8766	22,39533	229,382844		
16	Дисперсионный анализ						
17	Источник вариации	SS	df	MS	F	P-Значение	F критическое
18	Строки	713,9327	4	178,4832	0,88917132	0,613288688	3,259160053
19	Столбцы	225,2648	3	75,08819	0,2899356	0,831854274	3,490299605
20	Погрешность	3107,788	12	258,9823			
22	Итого	4046,985	19				

Рис. 5.36. Выходные результаты работы средства Двухфакторный дисперсионный анализ без повторений

В столбце Р-Значение вычисляются вероятности $P(X > F)$, где X — случайная величина, имеющая χ^2 -распределение со степенями свободы, значения которых приведены в столбце df: первое значение степени свободы — из соответствующей строки этого столбца, а второе — всегда из третьей строки, F — значение из столбца F. Например, значение в ячейке E18 (см. рис. 5.36) можно вычислить по формуле Excel =FPACn(D18;C18;C20). Эти значения используются для проверки гипотез о значимом влиянии факторов: если вероятность больше заданного уровня значимости, то нулевая гипотеза об отсутствии влияния принимается, в противном случае — отвергается.

В столбце F критическое вычисляются критические значения, соответствующие заданному в диалоговом окне Двухфакторный дисперсионный анализ без повторений уровню значимости α . Эти значения вычисляются как квантили порядка $1 - \alpha$ -распределения со степенями свободы, значения которых определяются так же, как при вычислении вероятностей из столбца Р-Значение. Например, значение в ячейке G18 (см. рис. 5.36) можно вычислить по формуле Excel =FPACnOBP(0,05;C18;C20). Эти значения используются для проверки гипотез о значимом влиянии факторов или их взаимного влияния: если значение в этом столбце больше значения в столбце F той же строки, то нулевая гипотеза об отсутствии влияния принимается, в противном случае — отвергается. Здесь принимаются обе нулевые гипотезы об отсутствии влияния факторов Р и у.

5.14. Корреляция

Это средство вычисляет корреляционную матрицу компонентов многомерной выборки. Диагональные элементы матрицы равны единице, а внедиагональные — коэффициентам корреляции соответствующих компонентов (о коэффициентах корреляции речь идет в разделе 1.2.5). На рис. 5.37 показаны многомерная выборка, имеющая совместное нормальное распределение, причем первая пара компонентов зависима с коэффициентом корреляции 0,5. С таким же коэффициентом корреляции зависимы третий и четвертый компоненты выборки. Первая и вторая пара компонентов между собой независимы.

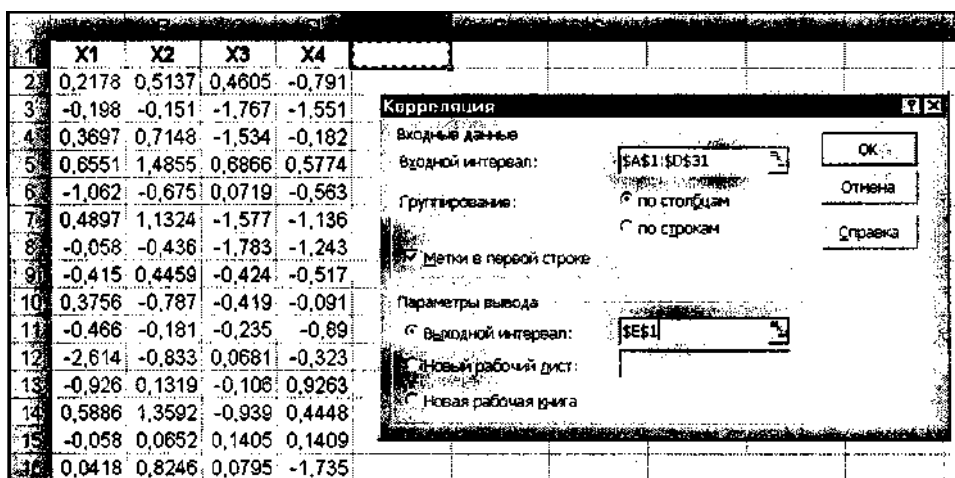


Рис. 5.37. Исходные данные и диалоговое окно Корреляция

Внедиагональные элементы корреляционной матрицы рассчитываются по стандартным формулам: коэффициент корреляции r_{xy} между компонентами x и y многомерной выборки вычисляется как

$$r_{xy} = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}, \text{ где } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, n - \text{объем выборки.}$$

Отметим, что эти же вычисления выполняет функция КОРРЕЛ (см. раздел 4.10.2).

На рис. 5.38 показан результат применения средства Корреляция. Поскольку корреляционная матрица симметрична, выводится только нижняя ее половина.

	A	B	C	D	E	F	G	H	
1	X1	X2	X3	X4		X1	X2	X3	X4
2	0,2178	0,5137	0,4605	-0,791	X1	1			
3	-0,198	-0,151	-1,767	-1,551	X2	0,539796	1		
4	0,3697	0,7148	-1,534	-0,182	X3	-0,11155	-0,05404	1	
5	0,6551	1,4855	0,6866	0,5774	X4	0,07675	0,079159	0,49465	1
6	-1,062	-0,675	0,0719	-0,563					
7	0,4897	1,1324	-1,577	-1,136					
8	-0,058	-0,436	-1,783	-1,243					
9	-0,415	0,4459	-0,424	-0,517					
10	0,3756	-0,787	-0,419	-0,091					
11	-0,466	-0,181	-0,235	-0,89					
12	-2,614	-0,833	0,0681	-0,323					

Рис. 5.38. Результат применения средства Корреляция

5.15. Ковариация

Это средство вычисляет ковариационную матрицу компонентов многомерной выборки. Диагональные элементы матрицы равны выборочным дисперсиям, а внедиагональные — ковариациям соответствующих компонентов (о ковариациях речь идет в разделе 1.2.5). На рис. 5.39 показана многомерная выборка, имеющая совместное нормальное распределение, причем первая пара компонентов зависима с коэффициентом корреляции 0,5. С таким же коэффициентом корреляции зависимы третий и четвертый компоненты выборки. Первая и вторая пары компонентов между собой независимы.

Внедиагональные элементы ковариационной матрицы рассчитываются по формулам: ковариация $\text{cov}(X, Y)$ между компонентами x и y многомерной выборки вычисляется как

$$\text{cov}(X, Y) = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}), \text{ где } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i, n - \text{объем выборки.}$$

Отметим, что эти же вычисления выполняет функция КОВАР (см. раздел 4.10.1). Диагональные элементы матрицы — выборочные дисперсии — вычисляются по

стандартным формулам $S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$. Выборочную дисперсию также вычисляют функции ДИСПР и ДИСПРА (см. раздел 4.5.2).

1	A	B	C	D	E	F	G	H	I	J
1	X1	X2	X3	X4						
2	0,2178	0,5137	0,4605	-0,791						
3	-0,198	-0,151	-1,767	-1,551						
4	0,3697	0,7148	-1,534	-0,182						
5	0,6551	1,4855	0,6866	0,5774						
6	-1,062	-0,675	0,0719	-0,563						
7	0,4897	1,1324	-1,577	-1,136						
8	-0,058	-0,436	-1,783	-1,243						
9	-0,415	0,4459	-0,424	-0,517						
10	0,3756	-0,787	-0,419	-0,091						
11	-0,466	-0,181	-0,235	-0,89						
12	-2,614	-0,833	0,0681	-0,323						
13	-0,926	0,1319	-0,106	0,9263						
14	0,5886	1,3592	-0,939	0,4448						
15	-0,058	0,0652	0,1405	0,1409						
16	0,0418	0,8246	0,0795	-1,735						

Ковариация

Входные данные: \$A\$1:\$D\$31

Группирование: ☐ по столбцам ☐ по строкам

☒ Метки в первой строке

Параметры вывода: Выходной интервал: \$E\$1

☐ Новый рабочий лист: ☐ Новая рабочая книга

ОК Отмена Справка

Рис. 5.39. Исходные данные и диалоговое окно Ковариация

На рис. 5.40 показан результат применения средства Ковариация. Поскольку ковариационная матрица симметрична, выводится только нижняя ее половина.

1	A	B	C	D	E	F	G	H	I
1	X1	X2	X3	X4		X1	X2	X3	X4
2	0,2178	0,5137	0,4605	-0,791	X1	0,888266			
3	-0,198	-0,151	-1,767	-1,551	X2	0,466787	0,841851		
4	0,3697	0,7148	-1,534	-0,182	X3	-0,10688	-0,05041	1,033594	
5	0,6551	1,4855	0,6866	0,5774	X4	0,062676	0,062931	0,435737	0,750764
6	-1,062	-0,675	0,0719	-0,563					
7	0,4897	1,1324	-1,577	-1,136					
8	-0,058	-0,436	-1,783	-1,243					
9	-0,415	0,4459	-0,424	-0,517					
10	0,3756	-0,787	-0,419	-0,091					
11	-0,466	-0,181	-0,235	-0,89					
12	-2,614	-0,833	0,0681	-0,323					

Рис. 5.40. Результат применения средства Ковариация

5.16. Регрессия

Задачи регрессионного анализа описаны в разделе 3.4. Покажем, что для проведения регрессионного анализа может сделать средство Регрессия. В отдельных таблицах оно вычисляет (рис. 5.42 и 5.43) следующее:

- методом наименьших квадратов — коэффициенты линейной (относительно этих коэффициентов) функции регрессии; вид функции регрессии определяется структурой исходных данных (подробнее об этом речь идет ниже);

- коэффициент детерминации и связанные с ним величины (таблица Регрессионная статистика);
- дисперсионную таблицу и критериальную статистику для проверки значимости регрессии (таблица Дисперсионный анализ);
- для каждого коэффициента регрессии — среднее квадратическое отклонение и другие его статистические характеристики, позволяющие проверить значимость этого коэффициента и построить для него доверительные интервалы;
- значения функции регрессии и *остатки* — разности между исходными значениями переменной Y и вычисленными значениями функции регрессии (таблица Вывод остатка);
- вероятности, соответствующие упорядоченным по возрастанию значениям переменной Y (таблица Вывод вероятности).

Кроме того, средство Регрессия строит три типа графиков, которые будут показаны ниже.

Пусть входной интервал X состоит из k диапазонов-столбцов, содержащих значения $\{x_{i1}\}, \{x_{i2}\}, \dots, \{x_{ik}\}$ переменных X_1, X_2, \dots, X_k . В каждом диапазоне содержится одинаковое количество значений. Входной интервал Y, состоящий из одного диапазона-столбца, должен содержать такое же количество значений. Средство Регрессия вычисляет коэффициенты функции регрессии вида

$$Y = m_1X_1 + m_2X_2 + \dots + m_kX_k + b.$$

Это уравнение линейной множественной регрессии, если переменные X_i независимы. На основе данного уравнения, используя соответствующие значения переменных X_i , можно получить множество других уравнений регрессии. Например, если в качестве переменных X_i взять значения одной переменной X в степени i (т.е. $X_i = X^i$), получим уравнение полиномиальной регрессии

$$Y = m_1X + m_2X^2 + \dots + m_kX^k + b.$$

На рис. 5.41 показан рабочий лист с исходными данными: входной интервал X состоит из пяти столбцов. В первом столбце представлены значения переменной X_1 , во втором — квадраты значений переменной X_1 , в третьем — значения второй переменной X_2 , в четвертом — квадраты значений переменной X_2 , в пятом — произведения значений переменных X_1 и X_2 . Таким образом, в данном случае Регрессия будет вычислять значения коэффициентов уравнения регрессии вида

$$Y = m_1X_1 + m_2X_1^2 + m_3X_2 + m_4X_2^2 + m_5X_1X_2 + b.$$

Отметим, что значения зависимой переменной Y в столбце F получены по формуле

$$Y = X_1 - 2X_1^2 + 0,5X_2 - X_2^2 + 5X_1X_2 + \epsilon.$$

Здесь случайная переменная ϵ имеет стандартное нормальное распределение. (О моделировании случайных величин речь идет в главе 7.)

Диалоговое окно средства Регрессия показано на рис. 5.41. В поле Входной интервал Y вводится адрес диапазона, содержащего значения зависимой переменной Y. Диапазон должен состоять из одного столбца. В поле Входной интервал X вводится адрес диапазона, содержащего значения переменной X. Диапазон должен состоять из одного или нескольких столбцов, но не более чем из 16 столбцов. Если указанные в полях Входной интервал Y и Входной

интервал X диапазоны включают заголовки столбцов, то необходимо установить флажок опции Метки — эти заголовки будут использованы в выходных таблицах, сгенерированных средством Регрессия.

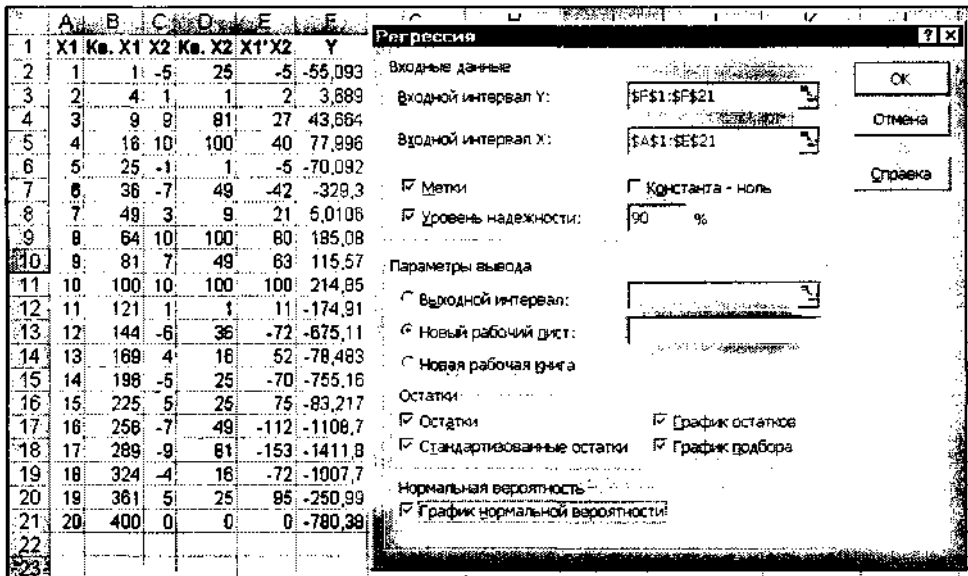


Рис. 5.41. Исходные данные и диалоговое окно Регрессия

Флажок опции Константа - ноль следует установить, если в уравнении регрессии константа B принудительно полагается равной нулю. Опция Уровень надежности устанавливается тогда, когда необходимо построить доверительные интервалы для коэффициентов регрессии с доверительным уровнем, отличным от 0,95, который используется по умолчанию. После установки флажка опции Уровень надежности становится доступным поле ввода, в котором вводится новое значение доверительного уровня.

В области Остатки имеются четыре опции: Остатки, Стандартизованные остатки, График остатков и График подбора. Если установлена хотя бы одна из них, то в выходных результатах появится таблица Вывод остатка, в которой будут выведены значения функции регрессии и остатки — разности между исходными значениями переменной Y и вычисленными значениями функции регрессии. Значения этой таблицы и возможности каждой из опций показаны ниже.

В области Нормальная вероятность имеется одна опция — График нормальной вероятности; ее установка порождает в выходных результатах таблицу Вывод вероятности и приводит к построению соответствующего графика.

На рис. 5.42-5.44 показаны части рабочего листа с выходными результатами средства Регрессия, которые получены на основе исходных данных, приведенных на рис. 5.41. Рассмотрим подробнее эти результаты.

В таблице Регрессионная статистика приводятся следующие данные.

- Множественный R — корень из коэффициента детерминации L^2 , приведенного в следующей строке. Другое название этого показателя — индекс корреляции, или множественный коэффициент корреляции (см. раздел 3.3.1).

	A	B	C	D	E	F
1	ВЫВОД ИТОГОВ					
2						
3	Регрессионная статистика					
4	Множественный R	0,999998881				
5	R-квадрат	0,999997761				
6	Нормированный R-квадрат	0,999996882				
7	Стандартная ошибка	0,83074245				
8	Наблюдения	20				
9						
10	Дисперсионный анализ					
11		<i>df</i>	<i>SS</i>	<i>MS</i>	<i>F</i>	<i>Значимость F</i>
12	Регрессия	5	4315679,352	863135,8704	1250680,44	5,01903E-39
13	Остаток	14	9,661862258	0,690133018		
14	Итого	19	4315689,014			

Рис. 5.42. Верхняя часть рабочего листа с выходными результатами

	A	B	C	D	E	F	G	H	I
16	Коэффициенты Стандартная ошибка t-статистика P-Значение Нижние 95% Верхние 95% Нижние 90% Верхние 90%								
17	У-пересечение	-0,253888671	0,648126025	-0,391696463	0,7011799	-1,643962	1,13622484	-1,395419	0,8875816
18	X1	1,123005378	0,138017336	8,136697854	1,124E-05	0,8269874	1,41902339	0,8799142	1,3660965
19	Кв. X1	-2,005986197	0,006477434	-308,6883928	2,956E-28	-2,0198789	-1,9920935	-2,017395	-1,984577
20	X2	0,672289116	0,083366283	8,064280793	1,241E-06	0,4934861	0,85108217	0,5254553	0,8191226
21	Кв. X2	-1,010435244	0,007081425	-143,0922648	1,457E-23	-1,0255805	-0,99529	-1,022873	-0,997997
22	X1^2	4,98589396	0,007028912	709,3407928	2,704E-33	4,9708184	5,00096949	4,9735139	4,9882740
23									
24	ВЫВОД ОСТАТКА				ВЫВОД ВЕРОЯТНОСТИ				
25	Наблюдение	Предсказанное Y	Остатки	Стандартные остатки	Перцентиль Y				
26	1	-54,68864596	-0,404791015	-0,567645558	2,5	-1411,7616			
27	2	3,601838089	0,087190777	0,12226915	7,5	-1108,7249			
28	3	43,88575591	-0,221342134	-0,310391986	12,5	-1007,8897			
29	4	77,25749888	0,738473754	1,03557473	17,5	-780,37807			
30	5	-71,40069086	1,30825897	1,834594557	22,5	-755,15699			
31	6	-329,3562365	0,051581852	0,072334139	27,5	-675,10639			
32	7	6,940566647	-1,929928977	-2,706373339	32,5	-329,30465			
33	8	184,8979413	0,184368329	0,258542949	37,5	-250,99418			
34	9	116,6743141	-1,106508782	-1,55167672	42,5	-174,81074			
35									
36	17	-1411,830421	-0,131188723	-0,183988253	82,5	77,9959728			
37	18	-1007,819785	0,15010449	0,210494165	87,5	115,567805			
38	19	-251,3172928	0,323114831	0,453109609	92,5	185,08231			
39	20	-780,1882398	-0,187832093	-0,263400247	97,5	214,845179			

Рис. 5.43. Нижняя часть рабочего листа с выходными результатами

- R-квадрат — коэффициент детерминации R^2 ; вычисляется как отношение регрессионной суммы квадратов (ячейка C12) к полной сумме квадратов (ячейка C14). (О коэффициенте детерминации речь идет в разделе 3.4.3.)
- Нормированный R-квадрат вычисляется по формуле $\frac{(n-1)R^2 - k}{n - k - 1}$, где n — количество значений переменной Y, k — количество столбцов во входном интервале переменной X.
- Стандартная ошибка — корень из остаточной дисперсии (ячейка D13).
- Наблюдения — количество значений переменной Y.

Дисперсионная таблица соответствует аналогичной таблице из раздела 3.4.3. В столбце SS приводятся суммы квадратов, в столбце df — число степеней свободы, в столбце MS — дисперсии. Строка Регрессия соответствует одноименной

строке из таблицы в разделе 3.4.3, строка Остаток — строке Остатки и строка Итого — строке Полная. В дисперсионной таблице из раздела 3.4.3 приведены формулы, по которым вычисляет соответствующие значения средство Регрессия. В столбце F вычислено значение критериальной статистики для проверки значимости регрессии. Это значение вычисляется как отношение регрессионной дисперсии к остаточной (ячейки D12 и D13). В столбце Значимость F вычисляется вероятность полученного значения критериальной статистики. (Эту вероятность с помощью формул Excel можно вычислить как =FPACn(E12;B12;B13).) Если эта вероятность меньше, например, 0,05 (заданного уровня значимости), то гипотеза о незначимости регрессии (т.е. гипотеза о том, что все коэффициенты функции регрессии равны нулю) отвергается и считается, что регрессия значима. В данном примере регрессия значима практически с любым уровнем значимости.

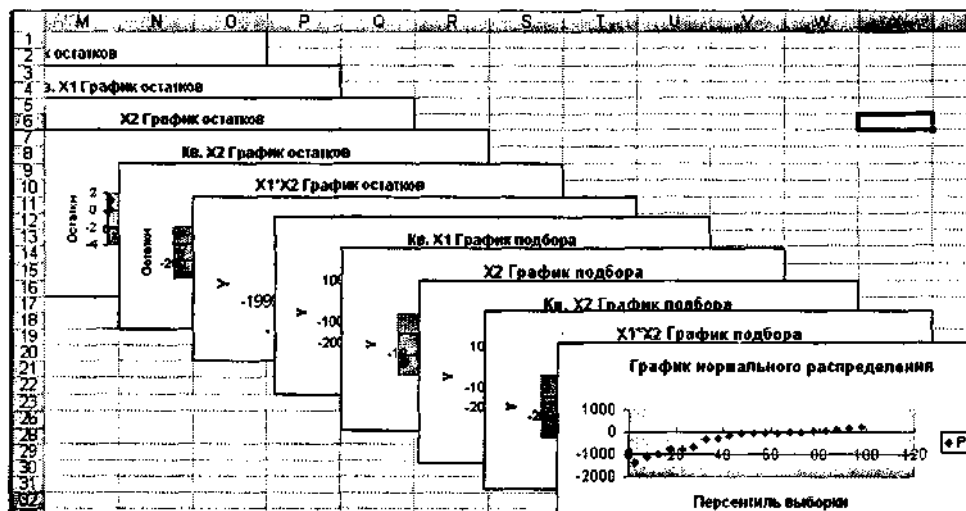


Рис. 5.44. Графики, выводимые средством Регрессия

В следующей таблице (см. рис. 5.43), в столбце Коэффициенты, записаны вычисленные значения коэффициентов функции регрессии, при этом в строке Y-пересечение записано значение свободного члена b_0 . В столбце Стандартная ошибка вычислены среднеквадратические отклонения коэффициентов (о вычислении дисперсий коэффициентов речь идет в разделе 3.4.4). В столбце t-статистика записаны отношения значений коэффициентов к их среднеквадратическим отклонениям. Это значения критериальных статистик для проверки гипотез о значимости коэффициентов регрессии. В столбце P-Значение вычисляются уровни значимости, соответствующие значениям критериальных статистик. (Их можно вычислить с помощью формулы Excel =СТБОПРАСН(ABS(D17);14;2), например, для значения в ячейке E17; второй аргумент в функции СТБЮДРАСП вычисляется как $n - k - 1$.) Если вычисленный уровень значимости меньше заданного уровня значимости (например, 0,05), то принимается гипотеза о значимом отличии коэффициента от нуля; в противном случае принимается гипотеза о незначимом отличии коэффициента от нуля. В данном примере только коэффициент b незначимо отличается от нуля.

В столбцах Нижние 95% и Верхние 95% приводятся границы доверительных интервалов с доверительным уровнем 0,95. Эти границы вычисляются по формулам

$$\text{Нижние 95\%} = \text{Коэффициент} - \text{Стандартная ошибка} \times t_{\alpha};$$

$$\text{Верхние 95\%} = \text{Коэффициент} + \text{Стандартная ошибка} \times t_{\alpha}.$$

Здесь t_{α} — квантиль порядка α распределения Стьюдента с $(n - k - 1)$ степенью свободы. В данном случае $\alpha = 0,95$. Аналогично вычисляются границы доверительных интервалов в столбцах Нижние 90,0% и Верхние 90,0%. Отметим, что если в диалоговом окне Регрессия не устанавливать опцию Уровень надежности, то будут повторены столбцы Нижние 95% и Верхние 95%.

Рассмотрим таблицу Вывод остатка из выходных результатов средства Регрессия. Напомним, что эта таблица появляется в выходных результатах только тогда, когда установлена хотя бы одна опция в области Остатки диалогового окна Регрессия. В столбце Наблюдение приводятся порядковые номера значений переменной Y . В столбце Предсказанное Y вычисляются значения функции регрессии $\hat{y}_i = f(x_i)$ для тех значений переменной X , которым соответствует порядковый номер i в столбце Наблюдение. В столбце Остатки содержатся разности (остатки) $\varepsilon_i = y_i - \hat{y}_i$, а в столбце Стандартные остатки — нормированные остатки, которые вычисляются как отношения $\varepsilon_i / s_{\varepsilon}$, где s_{ε} — среднеквадратическое отклонение остатков. Квадрат величины s_{ε} вычисляется по формуле

$$s_{\varepsilon}^2 = \frac{1}{n-1} \sum_{i=1}^n (\varepsilon_i - \bar{\varepsilon})^2, \text{ где } \bar{\varepsilon} — \text{среднее остатков. Здесь величину } s_{\varepsilon}^2 \text{ можно вычис-$$

лить как отношение двух значений из дисперсионной таблицы: суммы квадратов остатков (ячейка C13) и степени свободы из строки Итого (ячейка B14).

По значениям таблицы Вывод остатка средство Регрессия строит два типа графиков: графики остатков и графики подбора (если установлены соответствующие опции в области Остатки диалогового окна Регрессия). На рис. 5.45 показаны образцы этих графиков (графики немного переформатированы по сравнению с оригиналами). Они строятся для каждого компонента переменной X в отдельности. На графиках остатков отображаются остатки, т.е. разности между исходными значениями Y и вычисленными по функции регрессии для каждого значения компонента переменной X . На графиках подбора отображаются как исходные значения Y , так и вычисленные значения функции регрессии для каждого значения компонента переменной X . (На графиках подбора, представленных на рис. 5.45, эти значения практически совпадают.)

Последней таблицей выходных результатов средства Регрессия является таблица Вывод вероятности (см. рис. 5.43). Она появляется, если в диалоговом окне Регрессия установлена опция График нормальной вероятности. Значения в столбце Персентиль вычисляются следующим образом. Вычисляется шаг $h = (1/n) \times 100\%$, первое значение равно $h/2$, последнее равно $100 - h/2$. Начиная со второго значения каждое последующее значение равно предыдущему, к которому прибавлен шаг h . В столбце Y приведены значения переменной Y , упорядоченные по возрастанию. По данным этой таблицы строится так называемый график нормального распределения (рис. 5.46). Он позволяет визуально оценить степень линейности зависимости между переменными X и Y .

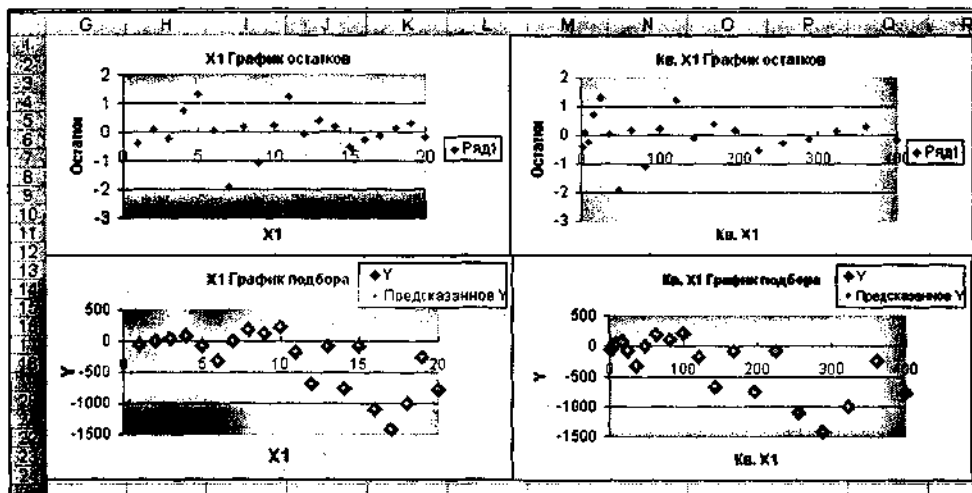


Рис. 5.45. Примеры графиков остатков и подбора

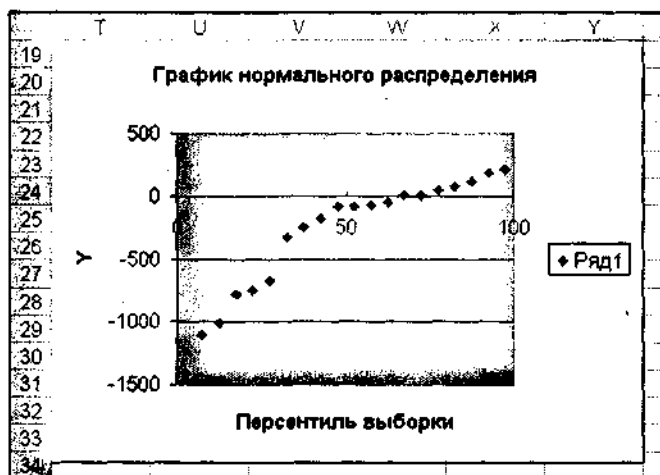


Рис. 5.46. График нормального распределения

5.17. Скользящее среднее

Метод скользящего среднего — один из наиболее широко используемых способов сглаживания значений временного ряда. Метод основан на локальном усреднении, когда за новое значение временного ряда берется среднее k последовательных значений, ближайших к заменяемому значению.

Пусть имеются дискретные наблюдения y_1, y_2, \dots, y_n , и задано число k наблюдений, по которым будет проводиться усреднение. Значение скользящего среднего для значения t вычисляется по формуле $y_t = \frac{1}{k} \sum_{i=0}^{k-1} y_{t-i}$. Отметим, что по этой

формуле выполняет вычисления средство Скользящее среднее, но существуют и другие способы вычисления скользящего среднего.

На рис. 5.47 показаны исходные данные, для которых будут вычисляться скользящие средние, и диалоговое окно Скользящее среднее. В поле ввода Входной интервал в качестве исходных данных задан диапазон B1:B17. Поскольку этот диапазон содержит заголовок, установлен флажок опции Метки в первой строке. В поле Интервал вводится число k — количество значений, по которым подсчитывается скользящее среднее. Если этот параметр не задан, то по умолчанию используется значение 3.

Если установлен флажок опций Вывод графика, то будет построен график, отображающий исходные значения y_t и сглаженные скользящим средним значения (рис. 5.48). Если также установлен флажок опции Стандартные погрешности, то к значениям вычисленных средних будет добавлен столбец, в котором будут записаны стандартные погрешности, вычисляемые как сумма квадратов разностей между исходными и расчетными k значения y_t деленная на число k . Формула Excel, по которой подсчитываются стандартные погрешности, показана на рис. 5.48.

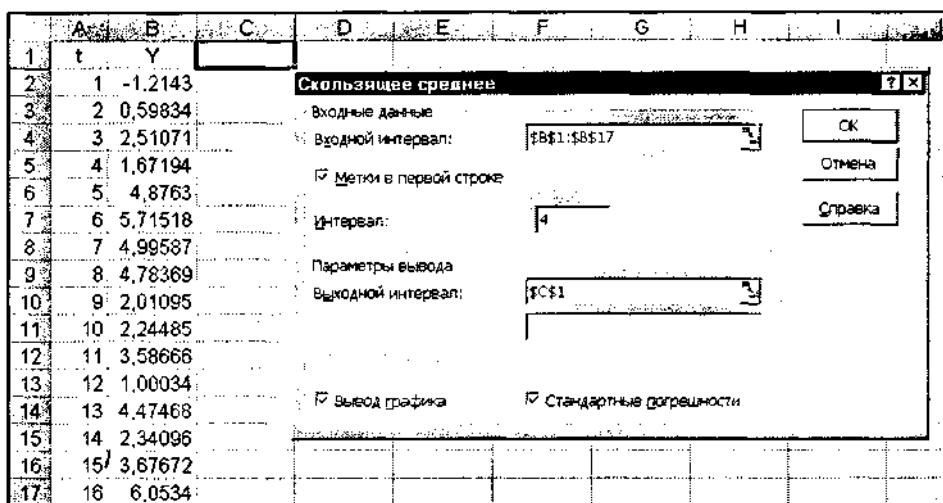


Рис. 5.47. Исходные данные и диалоговое окно Скользящее среднее

5.18. Экспоненциальное сглаживание

Экспоненциальное сглаживание, как и скользящее среднее (см. раздел 5.17), используется для выравнивания (сглаживания) значений временных рядов. Если имеются дискретные наблюдения y_1, y_2, \dots, y_n , то сглаженные значения вычисляются по формуле $\hat{y}_{t+1} = \alpha y_t + (1 - \alpha) \hat{y}_t$, где \hat{y}_t — сглаженное значение для предыдущего t , α — постоянная сглаживания, также называемая фактором затухания (это число из интервала $(0, 1)$).

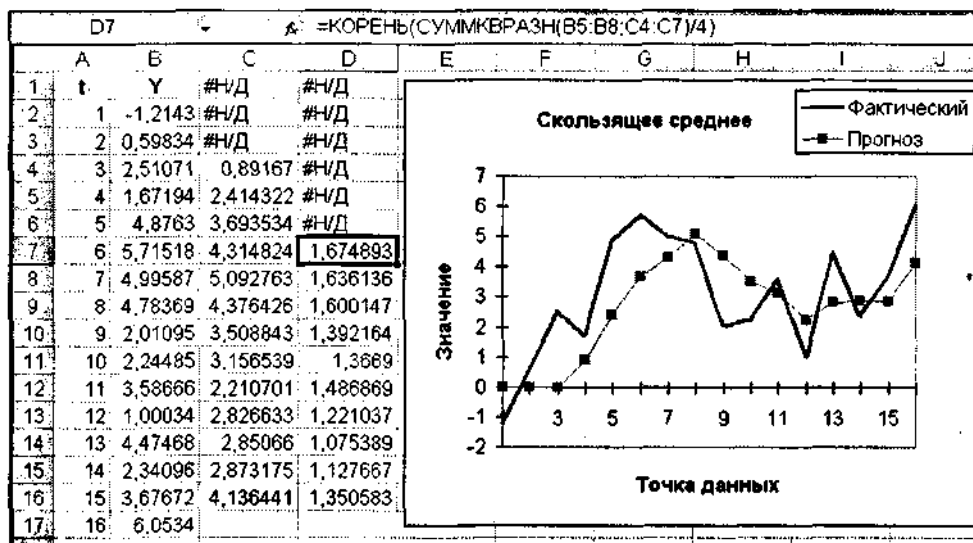


Рис. 5.48. Результаты вычислений

На рис. 5.49 показаны рабочий лист Excel с исходными данными (данные взяты из примера предыдущего раздела) и диалоговое окно Экспоненциальное сглаживание. В поле Входной интервал указывается адрес диапазона, содержащего значения y_t . Если этот диапазон включает заголовок, то надо установить флажок опции Метки. В поле Фактор затухания задается постоянная сглаживания; если она не задана, то по умолчанию используется значение 0,3. Установка флажков опций Вывод графика и Стандартные погрешности приводит к построению графика, на котором будут отображаться исходные и сглаженные значения (рис. 5.50), и к выводу дополнительного столбца со значениями погрешностей. Эти погрешности вычисляются как сумма квадратов разностей между тремя последовательными исходными и расчетными значениями, деленная на число 3. Формула Excel, по которой подсчитываются стандартные погрешности, показана на рис. 5.50.

5.19. Анализ Фурье

Данное средство выполняет дискретное преобразование Фурье. Это преобразование используется в анализе линейных систем и применяется к временным рядам для выявления периодических (спектральных) составляющих таких рядов.

Если имеются дискретные наблюдения y_1, y_2, \dots, y_n , то прямое дискретное преобразование Фурье выполняется в соответствии с формулой $Y_k = \sum_{j=1}^n y_j e^{-i \frac{2\pi}{n} jk}$, $k = 0,$

$1, \dots, n-1$. Результаты преобразования Y_k являются комплексными числами, модуль которых равен амплитуде k -й спектральной составляющей (k -й гармоники), а аргумент комплексного числа Y_k равен фазе этой гармоники. Аналогично опре-

деляется обратное дискретное преобразование Фурье ($y_j = \sum_{k=0}^{n-1} Y_k e^{i \frac{2\pi}{n} jk}$), которое преобразует спектральное представление временного ряда в действительное.

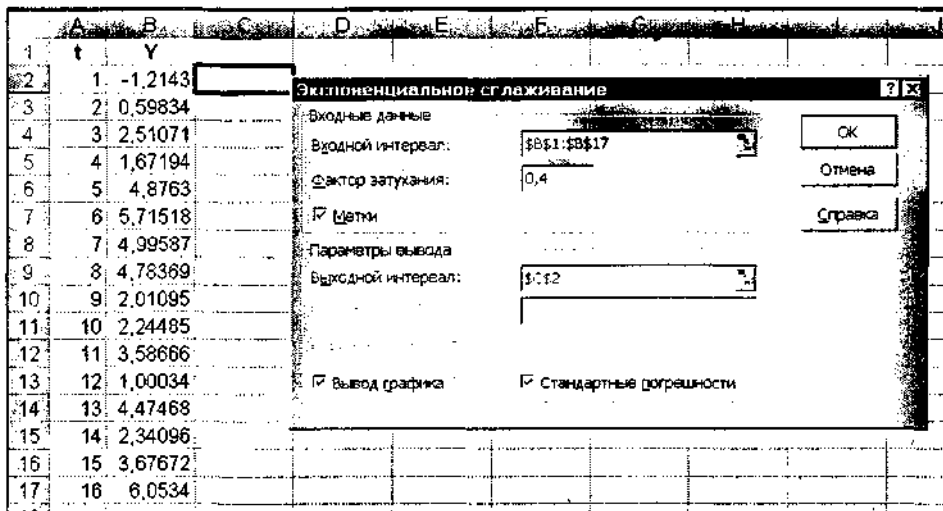


Рис. 5.49. Исходные данные и диалоговое окно Экспоненциальное сглаживание

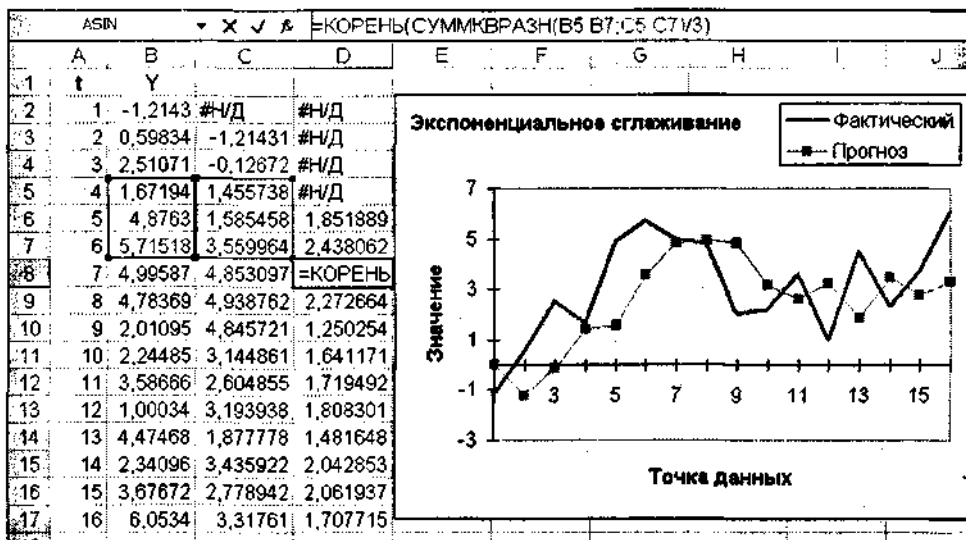


Рис. 5.50. Результаты вычислений

Средство Анализ Фурье выполняет как прямое так и обратное преобразования методом быстрого преобразования Фурье (БПФ). Применение метода БПФ диктует условие, чтобы количество исходных значений как для прямого, так и для обратного преобразований, было равно некоторой положительной степени числа 2. Максимальное число значений, которое может обработать средство Анализ Фурье, составляет 4096 ($= 2^{12}$). Для применения обратного преобразования Фурье исходные значения должны быть в формате комплексных чисел $x + yi$ или $x + yj$ (i и j — обозначение мнимой единицы). Если x является отрицательным числом, перед ним ставится апостроф (').

На рис. 5.51 показаны рабочий лист с исходными данными и диалоговое окно Анализ Фурье. Результат прямого преобразования Фурье показан на рис. 5.52. Первое значение (ячейка C2) равно сумме исходных данных.

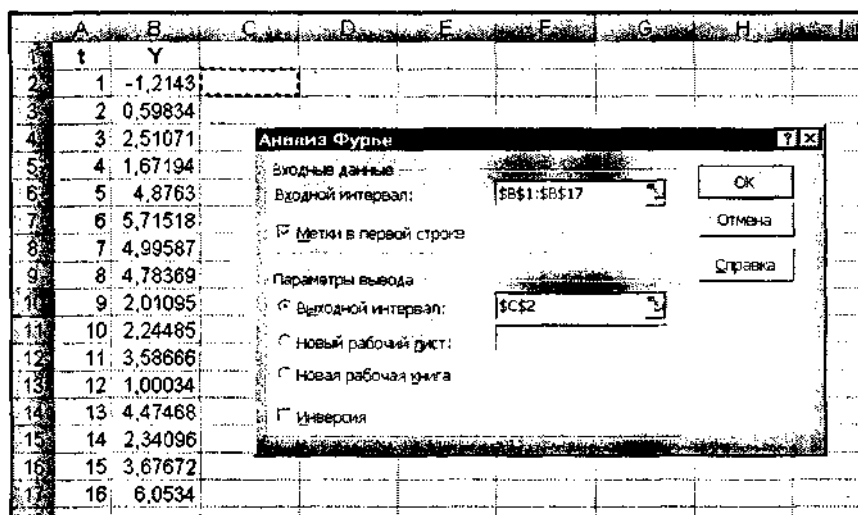


Рис. 5.51. Исходные данные и диалоговое окно Анализ Фурье

t	Y	Y комплексное
1	-1,2143	49,3262968701278
2	0,59834	-6,30122668866927-3,1954561628609i
3	2,51071	-6,46705511956092+12,0347360926031i
4	1,67194	0,821035859520582+4,47215620909899i
5	4,8763	-4,6223323514806+2,61004208030321i
6	5,71518	-3,88437851609255+4,01286152893839i
7	4,99587	-10,6416229921968+6,88428999446164i
8	4,78369	-3,53648384452208-2,04828062788192i
9	2,01095	0,508894946684288
10	2,24485	-3,53648384452207+2,04828062788193i
11	3,58666	-10,6416229921968-6,88428999446163i
12	1,00034	-3,88437851609256-4,01286152893838i
13	4,47468	-4,6223323514806-2,61004208030321i
14	2,34096	0,82103585952057-4,472156209099i
15	3,67672	-6,46705511956095-12,0347360926031i
16	6,0534	-6,30122668866926+3,19545616286091i

Рис. 5.52. Результат прямого преобразования Фурье

На рис. 5.53 показаны рабочий лист с исходными данными (результат прямого преобразования Фурье) для обратного преобразования и диалоговое окно Анализ Фурье, в котором установлен флажок опции Инверсия. Результат обратного преобразования Фурье показан на рис. 5.54; он совпадает с первоначальными данными из столбца B.

	A	B	C	D	E	F	G
1	t	Y	Y комплексное				
2	1	-1,2143	49,3262968701278				
3	2	0,59834	-6,30122668866927-3,1954561628609i				
4	3	2,51071	-6,46705511956092+12,0347360926031i				
5	4	1,67194	0,821035859520582+4,47215620909899i				
6	5	4,8763	-4,6223323514806+	<div> <div>Анализ Фурье</div> <div> <div>Входные данные</div> <div> Входной интервал: <input type="text" value="\$A\$1:\$F\$17"/> <div>OK</div> <div>Отмена</div> <div>Справка</div> </div> </div> <div> <div>Метки в первой строке</div> <div> <input checked="" type="checkbox"/> </div> </div> <div> <div>Параметры вывода</div> <div> <div>Выходной интервал: <input type="text" value="\$D\$2"/></div> <div> <input type="checkbox"/> Новый рабочий лист: <input type="text"/> <input type="checkbox"/> Новая рабочая книга </div> </div> <div> <input checked="" type="checkbox"/> Инверсия </div> </div> </div>			
7	6	5,71518	-3,88437851609255				
8	7	4,99587	-10,6416229921968				
9	8	4,78369	-3,53648384452208				
10	9	2,01095	0,508894946664288				
11	10	2,24485	-3,53648384452207				
12	11	3,58666	-10,6416229921968				
13	12	1,00034	-3,88437851609256				
14	13	4,47468	-4,6223323514806-				
15	14	2,34096	0,82103585952057-				
16	15	3,67672	-6,46705511956095				
17	16	6,0534	-6,30122668866926				
18							

Рис. 5.53. Исходные данные для обратного преобразования Фурье и диалоговое окно Анализ Фурье

	A	B	C	D
1	t	Y	Y комплексное	
2	1	-1,2143	49,3262968701278	-1,2143084680757
3	2	0,59834	-6,30122668866927-3,1954561628609i	0,598338032876482
4	3	2,51071	-6,46705511956092+12,0347360926031i	2,51071073937697
5	4	1,67194	0,821035859520582+4,47215620909899i	1,67193820535726
6	5	4,8763	-4,6223323514806+2,61004208030321i	4,8763014854766
7	6	5,71518	-3,88437851609255+4,01286152893839i	5,71518492388823
8	7	4,99587	-10,6416229921968+6,88428999446164i	4,99587262986481
9	8	4,78369	-3,53648384452208-2,04828062788192i	4,78369276345569
10	9	2,01095	0,508894946664288	2,01095482936513
11	10	2,24485	-3,53648384452207+2,04828062788193i	2,24485009039457
12	11	3,58666	-10,6416229921968-6,88428999446163i	3,58665980105682
13	12	1,00034	-3,88437851609256-4,01286152893838i	1,00033833243323
14	13	4,47468	-4,6223323514806-2,61004208030321i	4,4746839316917
15	14	2,34096	0,82103585952057-4,472156209099i	2,34095639355499
16	15	3,67672	-6,46705511956095-12,0347360926031i	3,67672095963972
17	16	6,0534	-6,30122668866926+3,19545616286091i	6,0534022197713
18				

Рис. 5.54. Результат обратного преобразования Фурье

Дополнительные возможности Excel для проведения статистического анализа

В этой главе описаны средства Excel общего назначения, которые не всегда рассматриваются в "стандартном" учебном курсе по электронным таблицам либо рассматриваются недостаточно полно. Здесь приведены формулы массивов — мощное средство для проведения вычислений, некоторые возможности построения диаграмм, полезные для визуализации статистических данных, надстройка Поиск решения — средство для решения оптимизационных задач, которое можно применить и при проведении статистического анализа, а также другие возможности Excel.

6.1. Массивы и формулы массивов

В этом разделе рассмотрены два основных понятия, которые зачастую значительно упрощают проведение вычислений в Excel. Это массив и формула массива. *Массив* — набор ячеек или значений, которые обрабатываются как единая группа. Элементы массива могут содержаться в группе ячеек или быть *поименованной константой* (см. далее). *Формула массива* — формула, в которой используется один или несколько массивов непосредственно или в качестве аргументов функций и которая возвращает одно или несколько значений. Напомним, что некоторые статистические функции рационально использовать именно в виде формул массивов, например функцию РАНГ (см. раздел 4.2.5).

Итак, массив — это некоторый поименованный набор элементов. В Excel массивы могут быть одно- или двумерными. Одномерный массив может быть группой ячеек, которые размещены в одной строке (горизонтальный массив) или в одном столбце (вертикальный массив). Двумерный массив размещается в нескольких строках и столбцах. Отметим, что в *массивах констант* нельзя использовать ссылки на ячейки, имена диапазонов или формулы, но можно использовать текстовые значения, заключенные в кавычки, и логические значения ИСТИНА и ЛОЖЬ.

Операции над массивами производятся с помощью *формул массивов*. Чтобы создать формулу массива, выполните следующие действия.

1. Выделите ячейку (если формула массива возвращает только одно значение) или диапазон ячеек (если формула массива возвращает несколько значений).

2. Введите формулу.
3. Нажмите комбинацию клавиш <Ctrl+Shift+Enter>.

Excel поместит формулу массива во все выделенные ячейки и автоматически заключит формулы в фигурные скобки, чтобы подчеркнуть, что это формулы массива.

Рассмотрим пример. В парном тесте Стьюдента необходимо вычислить средние значения разностей парных наблюдений (об этом тесте речь идет в разделе 2.4.2).

	А	В	С
1	Парные наблюдения		
2	-0,3002	-0,1677	
3	-1,2777	-1,2075	
4	0,2443	0,1501	
5	1,2765	0,1491	
6	1,1984	0,5863	
7	1,7331	0,7365	
8	-2,1836	-1,3002	
9	-0,2342	-0,2022	
10	1,0950	0,7376	
11	-1,0867	-0,8437	
12	-0,6902	-0,5218	
13	-1,6904	-0,9949	
14	-1,8469	-1,1203	
15	-0,9776	-0,4635	
16	-0,7735	-0,3964	
17	-2,1179	-1,8147	
18	-0,5679	-0,5666	
19	-0,4040	-0,0831	
20	0,1349	0,0433	
21	-0,3655	-0,2023	

Рис. 6.1. Исходные данные

Пусть двумерный массив выборочных значений располагается в столбцах А и В, как показано на рис. 6.1. В столбце С будут выведены разности. Для их вычисления можно применить формулу =A2-B2, которая записывается в ячейке С2, и затем скопировать ее вниз на диапазон С3:С51 (имеется 50 наблюдений). В результате будет получен диапазон ячеек, содержащий разности парных наблюдений. То же самое можно сделать с помощью формулы массива. Выделите диапазон С2:С51, введите формулу =A2:A51-B2:B51 (рис. 6.2) и нажмите клавиши <Ctrl+Shift+Enter>. Результат показан на рис. 6.3.

Еще раз подчеркнем, что формула массива вводится путем нажатия клавиш <Ctrl+Shift+Enter>. Excel автоматически заключает формулы в фигурные скобки — вручную их вводить нельзя, это будет ошибкой и Excel не примет такую формулу.

Пока преимуществ формул массивов по сравнению с обычными формулами не видно (за исключением, возможно, времени, сэкономленного на копировании формулы). Теперь вычислим среднее этих разностей, для чего воспользуемся стандартной функцией СРЗНАЧ, как показано на рис. 6.4. С помощью формулы массива это же значение можно получить, не используя вычисленные разности! Для этого следует применить формулу =СРЗНАЧ(A2:A51-B2:B51), которая вводится только в одну ячейку (а не в диапазон ячеек), но по завершении ее ввода все равно необходимо нажать комбинацию клавиш <Ctrl+Shift+Enter>. Результат применения этой формулы показан на рис. 6.5.

Последняя формула уже демонстрирует преимущества формул массивов, поскольку она исключила необходимость выполнять промежуточные вычисления для нахождения разностей. В этой формуле используются два массива. Она вычисляет разности пар значений ячеек диапазонов А2:А51 и В2:В51 и создает в памяти компьютера новый временный массив, в который записывается результат попарных вычитаний. Функция СРЗНАЧ вычисляет среднее значение элементов нового массива и отображает его в ячейке. В сущности, формула выполнила циклические вычисления, которые затруднительно напрямую реализовать на рабочем листе Excel. Приведенные ниже примеры покажут другие достоинства формул массивов.

	A	B	C	D
1	Парные наблюдения		Разности	
2	-0,3002	-0,1677	=A2-A51-B2:B51	
3	-1,2777	-1,2075		
4	0,2443	0,1501		
5	1,2765	0,1491		
6	1,1984	0,5863		
7	1,7331	0,7365		
8	-2,1836	-1,3002		
9	-0,2342	-0,2022		
10	1,0950	0,7376		
11	-1,0867	-0,8437		
12	-0,6902	-0,5218		
13	-1,6904	-0,9949		
14	-1,8469	-1,1203		
15	-0,9776	-0,4635		
16	-0,7735	-0,3964		
17	-2,1179	-1,8147		
18	-0,5679	-0,5666		
19	-0,4040	-0,0831		
20	0,1349	0,0433		

Рис. 6.2. Создание формулы массива

	A	B	C	D
1	Парные наблюдения		Разности	
2	-0,3002	-0,1677	-0,13252	
3	-1,2777	-1,2075	-0,0702	
4	0,2443	0,1501	0,094139	
5	1,2765	0,1491	1,127357	
6	1,1984	0,5863	0,612022	
7	1,7331	0,7365	0,996647	
8	-2,1836	-1,3002	-0,88343	
9	-0,2342	-0,2022	-0,03199	
10	1,0950	0,7376	0,3574	
11	-1,0867	-0,8437	-0,24304	
12	-0,6902	-0,5218	-0,16835	
13	-1,6904	-0,9949	-0,69549	
14	-1,8469	-1,1203	-0,7266	
15	-0,9776	-0,4635	-0,51412	
16	-0,7735	-0,3964	-0,37713	
17	-2,1179	-1,8147	-0,30325	
18	-0,5679	-0,5666	-0,00133	
19	-0,4040	-0,0831	-0,32094	
20	0,1349	0,0433	0,091595	

Рис. 6.3. Вычисление формулы массива

	A	B	C	D	E
1	Парные наблюдения		Разности		
2	-0,3002	-0,1677	-0,13252	Среднее разностей=	0,007593
3	-1,2777	-1,2075	-0,0702		
4	0,2443	0,1501	0,094139		
5	1,2765	0,1491	1,127357		
6	1,1984	0,5863	0,612022		
7	1,7331	0,7365	0,996647		
8	-2,1836	-1,3002	-0,88343		
9	-0,2342	-0,2022	-0,03199		
10	1,0950	0,7376	0,3574		

Рис. 6.4. Результат вычисления обычной формулы

	A	B	C	D	E
1	Парные наблюдения		Разности		
2	-0,3002	-0,1677	-0,13252	Среднее разностей=	0,007593
3	-1,2777	-1,2075	-0,0702		
4	0,2443	0,1501	0,094139	Среднее разностей=	0,007593
5	1,2765	0,1491	1,127357		
6	1,1984	0,5863	0,612022		
7	1,7331	0,7365	0,996647		
8	-2,1836	-1,3002	-0,88343		
9	-0,2342	-0,2022	-0,03199		
10	1,0950	0,7376	0,3574		

Рис. 6.5. Результат вычисления формулы массива

6.1.1. Редактирование формул массивов

Сделаем несколько общих замечаний о редактировании формул массивов. Если формула массива помещена в несколько ячеек, следует редактировать все ячейки диапазона как одну ячейку, поскольку нельзя изменить только один элемент, содержащий формулу массива. Если попытаться сделать это, Excel выдаст окно с сообщением *Нельзя изменить часть массива*. Чтобы отредактировать формулу массива, выделите все ячейки массива и активизируйте строку формул (щелкните на ней или нажмите <F2>). При редактировании формулы Excel удаляет фигурные скобки. Закончив редактирование формулы, нажмите <Ctrl-t-Shift+Enter>, чтобы внести изменения. Теперь содержимое всех ячеек массива будет соответствовать внесенным изменениям. *(Совет. Чтобы быстро выделить весь массив, перейдите к одной из ячеек диапазона массива и нажмите комбинацию клавиш <Ctrl+>/>, где клавиша </> — это клавиша на дополнительной цифровой клавиатуре.)*

С формулами массивов нельзя делать следующее.

- Изменять содержимое одной из ячеек, в которых находится формула массива.
- Перемещать отдельные ячейки, на которые распространяется формула массива (можно перемещать только все ячейки с формулой массива сразу).
- Удалять отдельные ячейки, на которые распространяется формула массива (можно удалить только весь массив целиком).
- Вставлять новые ячейки в массив; это относится также к вставке новых строк или столбцов, которые добавляют новые ячейки к массиву.

Нельзя изменить формулу массива в отдельной его ячейке; тем не менее, можно форматировать весь массив или отдельные его части.

6.1.2. Массивы констант

В приведенном выше примере в качестве массивов использовались диапазоны ячеек. Однако в формулах можно также использовать массив констант. Такой массив можно ввести непосредственно в формулу (например, как аргумент функции) или определить заранее с помощью диалогового окна *Присвоение имени*. Массивы констант можно использовать в формулах вместо ссылки на диапазоны ячеек. Чтобы использовать массив констант, в формулу массива необходимо ввести набор значений и заключить его в фигурные скобки. Либо следует воспользоваться именем массива, если оно ему было предварительно присвоено.

Массив констант может быть как одно-, так и двумерным. Одномерные массивы могут быть вертикальными или горизонтальными. Элементы одномерного горизонтального массива отделяются один от другого точкой с запятой, например {1;2;3;4;5}. Элементы одномерного вертикального массива отделяются двоеточием. Например, вот как определяется шестизначный вертикальный массив: {1:2:3:4:5:6}. В двумерном массиве элементы одной строки также отделяются точкой с запятой, а строки отделяются одна от другой двоеточием. Вот пример массива размерностью 3x4 (три строки, каждая из которых занимает четыре столбца): {1;2;3;4;5;6;7;8;9;10;11;12}.

На рис. 6.6 показано, как можно создать поименованный массив констант с помощью диалогового окна *Присвоение имени* (это окно открывается с помощью команды *Вставка>Имя^Присвоить*).

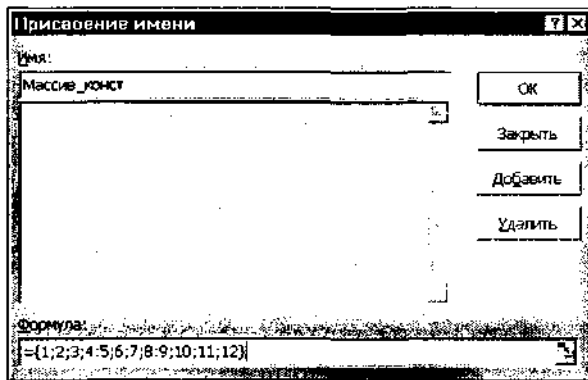


Рис. 6.6. Создание массива констант в диалоговом окне Присвоение имени

Способы использования массивов констант

Как говорилось выше, при задании функций массивы констант могут использоваться в качестве одного из аргументов, например `=СУММ(A1:A10;{1;3;5;7;})`. Здесь суммируются значения, содержащиеся в диапазоне A1:A10, и значения 1, 3, 5, 7. Вместо набора значений, заключенного в фигурные скобки, можно указать имя массива, если оно создано заранее. Пусть именем массива будет Массив_конст (о допустимых именах массивов сказано ниже), например `=СУММ(A1:A10;Массив_конст)`. Вводить такие функции, как формулы массивов, нет необходимости.

На рис. 6.7 приведено несколько простых примеров использования массивов констант непосредственно в формулах. Отметим, что все эти формулы вводились, как формулы массивов (предварительно был выделен необходимый диапазон ячеек), хотя заключительные фигурные скобки в режиме отображения формул Excel не показывает.

Чтобы перенести массив констант на рабочий лист Excel, выделите необходимый диапазон ячеек, введите формулу типа `= {1;3;5;7}` или `=Массив_конст` (если Массив_конст — имя существующего массива) и нажмите клавиши `<Ctrl-r-Shift+Enter>`. При этом следует помнить о размерности и ориентации массива. Например, массив `{1;3;5;7}` — горизонтальный массив (поскольку здесь для отделения элементов массива один от другого использованы точка с запятой), и если будет выделен вертикальный диапазон ячеек, то в этот диапазон будет записано только число 1 из этого массива.

6.1.3. Поименованные массивы и диапазоны

Использование в формулах адресов диапазонов ячеек очень утомительно и часто приводит к созданию формул, которые трудно читать и понимать (особенно через некоторое время). Excel позволяет присваивать ячейкам, диапазонам и массивам содержательные имена. Например, диапазон можно назвать Выборка или Стат_характеристики. Использование подобных имен (по сравнению с адресами ячеек или диапазонов) дает очевидные преимущества. Например, содержательное имя диапазона запомнить намного легче, чем адрес ячейки. Кроме того, при введении адресов ячеек и диапазонов легче ошибиться, чем при введении имен, а при выборе имени ячейки или диапазона это имя появляется в поле

Имя в строке формул. Применение имен значительно упрощает процесс создания формул — имя ячейки или диапазона можно вставить в формулу, используя команду Вставка^Имя^Вставить или выбрав соответствующее имя в поле Имя. Наконец, имена делают формулы более понятными и простыми в использовании.

B6

{=(1;2;3;4)-(4;3;2;1)}

Чар05_01.xls:2

	A	B	C	D	E	F
2	1	1,414214	1,732051	2		
3	2,236068	2,44949	2,645751	2,828427		
4	3	3,162278	3,316625	3,464102		
5						
6	1	-3	-1	1	3	
7	4					
8	9	-9	-8	-7		
9	16	-5	-4	-3		
10	25	-1	0	1		

Чар05_01.xls:3

	A	B	C	D	E
2	=МАССИВ_КОНСТ^(1/2)	=МАССИВ_КОНСТ^(1/2)	=МАССИВ_КОНСТ^(1/2)	=МАССИВ_КОНСТ^(1/2)	
3	=МАССИВ_КОНСТ^(1/2)	=МАССИВ_КОНСТ^(1/2)	=МАССИВ_КОНСТ^(1/2)	=МАССИВ_КОНСТ^(1/2)	
4	=МАССИВ_КОНСТ^(1/2)	=МАССИВ_КОНСТ^(1/2)	=МАССИВ_КОНСТ^(1/2)	=МАССИВ_КОНСТ^(1/2)	
5					
6	={1;2;3;4;5}^2	={1;2;3;4)-(4;3;2;1}	={1;2;3;4)-(4;3;2;1}	={1;2;3;4)-(4;3;2;1}	={1;2;3;4
7	={1;2;3;4;5}^2				
8	={1;2;3;4;5}^2	=МАССИВ_КОНСТ-10	=МАССИВ_КОНСТ-10	=МАССИВ_КОНСТ-10	
9	={1;2;3;4;5}^2	=МАССИВ_КОНСТ-10	=МАССИВ_КОНСТ-10	=МАССИВ_КОНСТ-10	
10	={1;2;3;4;5}^2	=МАССИВ_КОНСТ-10	=МАССИВ_КОНСТ-10	=МАССИВ_КОНСТ-10	

Рис. 6.7. Примеры использования массивов констант

Допустимые имена диапазонов и массивов

Хотя Excel достаточно "либеральна" в отношении имен диапазонов и массивов, существуют некоторые правила их выбора.

- В именах не должно быть никаких пробелов; для лучшего восприятия имени можете воспользоваться символом подчеркивания или точкой, например Среднее_выборки1 или Среднее.выборки1.
- Можно использовать любые комбинации букв и цифр, но имя не должно начинаться с цифры (например, 3-йРезультат) или быть похожим на адрес ячейки (например, A5).
- Имена должны содержать не больше 255 символов.
- В качестве имени можно использовать одиночные буквы (за исключением R и C), но это не рекомендуется делать, ведь смысл состоит именно в том, чтобы давать содержательные имена.

В Excel есть несколько имен для внутреннего употребления. И хотя можно создавать имена, замещающие внутренние имена Excel, лучше этого не делать. Поэтому следует избегать имен Область_печати, Заголовки_печати, Область_консолидации и Имя_листа.

Создание имен

Существует несколько способов создания имен.

Использование диалогового окна Присвоение имени. Чтобы создать имя для ячейки или диапазона, сначала выделите эту ячейку или диапазон. Затем выполните команду Вставка^ИмяОПрисвоить (или воспользуйтесь комбинацией клавиш <Ctrl+F3>). В результате Excel отобразит диалоговое окно Присвоение имени, показанное на рис. 6.6.

Введите имя в поле ввода Имя или воспользуйтесь именем, которое предложит программа (если она его, конечно, предложит). В текстовом поле Формула появится адрес активной или выбранной ячейки (или выделенного диапазона). Убедитесь в том, что это правильный адрес, а затем щелкните на кнопке ОК, чтобы добавить имя и закрыть диалоговое окно. Все введенные ранее имена отображаются в списке этого диалогового окна.

Использование поля Имя. Существует и более быстрый способ — создание имени с помощью поля Имя в строке формул. Выделите ячейку (или диапазон), которой нужно присвоить имя, а затем щелкните на этом поле и введите имя. Нажмите клавишу <Enter>, и имя будет создано. Поле Имя — это раскрывающийся список, в котором содержатся все имена, использующиеся в данной рабочей книге. Если выбрать ячейку (или диапазон), которой присвоено имя, это имя появится в поле Имя. Чтобы выбрать ячейку (или диапазон), которой присвоено имя, щелкните на поле Имя и выберите из списка нужное имя. В результате Excel выделит соответствующую ячейку или диапазон.

Автоматическое создание имен. Часто возникает необходимость использовать текст, содержащийся в рабочей таблице, для создания имен ячеек или диапазонов. На рис. 6.8 приведен пример такой таблицы. В данном случае может понадобиться использовать текст из ячеек A1 и B1 для создания имен соответствующих значений столбцов A и B. Excel позволяет это сделать легко и просто.

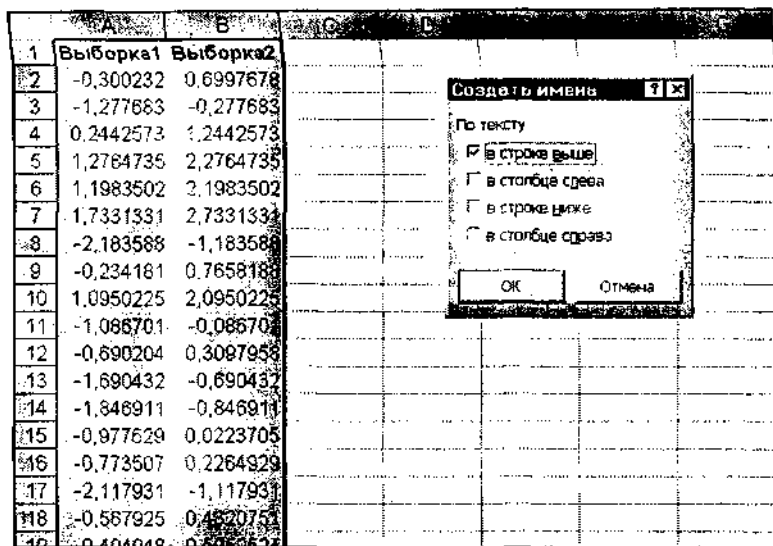


Рис. 6.8. Создание имен на основе текста, расположенного в соседних ячейках

Чтобы создать имена с помощью текста, расположенного в соседних ячейках, сначала выделите этот текст и ячейки, которые нужно назвать (это могут быть как отдельные ячейки, так и диапазоны). Затем выберите команду ВставкаО Имя=>Создать или нажмите комбинацию клавиш <Ctrl+Shift+F3>. В результате Excel отобразит диалоговое окно Создать имена, показанное на рис. 6.8. Флажок опции в этом диалоговом окне установлен на основании проведенного программой Excel анализа выделенного диапазона. Например, если программа обнаружила текст в первой строке выбранного множества ячеек, то она предложит создать имена на основе текста в верхней строке (опция в строке выше). Если догадка Excel неверна, вы можете выбрать другую опцию. Щелкните на кнопке ОК, и имена будут созданы.

Переопределение имен

После определения имени может понадобиться изменить ячейку (или диапазон), к которой оно относится. Для этого можно воспользоваться диалоговым окном Присвоение имени. Выберите команду Вставка^ИмяОПрисвоить, щелкните на имени, которое необходимо переопределить, и измените адрес ячейки или диапазона в поле редактирования Формула. То же самое можно сделать другим способом: щелкнуть на поле Формула и выбрать новую ячейку (или диапазон) в рабочем листе, указав на нее мышью. Excel автоматически исправляет адреса ячеек, имеющих имена.

6.1.4. Примеры использования формул массивов

В этом разделе представлено несколько примеров, которые демонстрируют использование формул массивов и поименованных диапазонов и массивов и которые могут быть полезными при проведении статистического анализа.

Использование условных выражений

Допустим, имеется одномерная выборка и необходимо подсчитать количество, сумму и среднее выборочных значений, которые больше выборочного среднего всей выборки, и аналогичные величины для значений, которые меньше выборочного среднего всей выборки. С помощью формул массивов такие вычисления выполняются относительно просто. Пусть диапазон выборочных значений имеет имя Данные. Формулы и результаты вычислений показаны на рис. 6.9.

К сожалению, такие вычисления нельзя выполнить с помощью функций СУММЕСЛИ и СЧЁТЕСЛИ, поскольку они не поддерживают в качестве аргументов, задающих условия отбора значений, формул массивов. Но их можно использовать с простыми (без формул) условиями отбора. Например, если надо подсчитать сумму только положительных значений из диапазона Данные, то вполне подойдет формула =СУММЕСЛИ(Данные;">0";Данные), введенная, как формула массива.

Суммирование k-х чисел в выборке

Следующий пример немного сложнее, чем предыдущий. Предположим, есть выборка и в ней необходимо вычислить сумму всех третьих чисел (в общем случае k-х чисел), т.е. сложить первое, четвертое, седьмое и т.д. числа, а также вычислить их среднее. Это можно сделать, предварительно применив средство Выборка из пакета анализа (см. раздел 5.4), которое выведет в отдельный массив эти третьи числа. Затем останется только подсчитать сумму и среднее чисел. Но

это же можно сделать с помощью формулы массива. Предположим, в ячейке с именем Период находится число k , диапазон выборочных значений имеет имя Выборка, выборочные значения пронумерованы и их номера находятся в диапазоне Номер, как показано на рис. 6.10.

D2		{=СЧЁТ(ЕСЛИ(Данные>=СРЗНАЧ(Данные),Данные,""))}								
	A	B	C	D	E	F	G	H	I	J
1	Данные	Значения, большие среднего выборки								
2	0	Количество	10	{=СЧЁТ(ЕСЛИ(Данные>=СРЗНАЧ(Данные),Данные,""))}						
3	9	Сумма	64	{=СУММ(ЕСЛИ(Данные>=СРЗНАЧ(Данные),Данные,""))}						
4	10	Среднее	6,4	{=СРЗНАЧ(ЕСЛИ(Данные>=СРЗНАЧ(Данные),Данные,""))}						
5	3									
6	6	Значения, меньшие среднего выборки								
7	3	Количество	9	{=СЧЁТ(ЕСЛИ(Данные<СРЗНАЧ(Данные),Данные,""))}						
8	2	Сумма	-14	{=СУММ(ЕСЛИ(Данные<СРЗНАЧ(Данные),Данные,""))}						
9	4	Среднее	-1,5556	{=СРЗНАЧ(ЕСЛИ(Данные<СРЗНАЧ(Данные),Данные,""))}						
10	-5									
11	-1									
12	-2									
13	-5									
14	6									
15	5									
16	-3									
17	0									
18	0									

Рис. 6.9. Использование условных выражений в формулах массивов

J2		{=ОСТАТ(Номер;Период)}								
	A	B	C	D	E	F	G	H	I	J
1	Номер	Выборка	Период		3	Функция ОСТАТ				
2	1	93								1
3	2	95	Сумма		212	{=ЕСЛИ(Период=0,0,СУММ(ЕСЛИ(ОСТАТ(Номер;Период)=0;Выборка,0)))}				
4	3	15								
5	4	30								
6	5	97	Среднее		10,6	{=ЕСЛИ(Период=0,0,СРЗНАЧ(ЕСЛИ(ОСТАТ(Номер;Период)=0;Выборка,0)))}				
7	6	4								
8	7	75								
9	8	52								
10	9	58								
11	10	83								
12	11	83								
13	12	89								
14	13	64								
15	14	6								
16	15	5								
17	16	1								
18	17	17								

Рис. 6.10. Для суммирования k -х выборочных значений и вычисления их среднего используются формулы массивов

Чтобы вычислить сумму k -х выборочных значений, используется формула массива

$\{\text{ЕСЛИ}(\text{Период}=0;0;\text{СУММ}(\text{ЕСЛИ}(\text{ОСТАТ}(\text{Номер};\text{Период})=0;\text{Выборка};0)))\}$.

Для определения значений элементов выборки, подлежащих суммированию, в формуле используется функция ОСТАТ. Она возвращает остаток от деления первого своего аргумента на второй аргумент. (Значения, возвращаемые этой функцией, показаны на рис. 6.10 в столбце J.) Если функция ОСТАТ возвращает 0, то число включается в массив суммирования. Обратите внимание, что случай, когда Период равен 0, рассмотрен отдельно, поскольку функция ОСТАТ возвращает ошибку, если ее второй аргумент равен 0.

Аналогично работает формула вычисления среднего

$\{=\text{ЕСЛИ}(\text{Период}=0;0;\text{СРЗНАЧ}(\text{ЕСЛИ}(\text{ОСТАТ}(\text{Номер};\text{Период})=0;\text{Выборка};0)))\}$.

Приведенные формулы используют массив номеров выборочных значений. Можно отказаться от этого массива и для отбора выборочных значений (точнее, для определения их последовательных номеров в выборке) применить функцию СТРОКА, которая возвращает номер строки, содержащей ее аргумент. Однако в этом случае формулы значительно усложняются, поскольку необходимо либо хранить адрес первой ячейки массива Выборка отдельно, либо находить его в процессе вычислений.

Вычисление рангов

В Excel для вычисления рангов выборочных значений существуют функция РАНГ (см. раздел 4.2.5) и средство Ранг и перцентиль (см. раздел 5.5). Способ, которым эти функция и средство устанавливают ранги, не всегда устраивает, поскольку одинаковым значениям они присваивают одинаковые ранги, равные рангу первого значения этой группы значений. Например, если есть два одинаковых значения, причем первому из них приписывается ранг, например, 5, тогда обоим значениям устанавливается тот же ранг 5. Однако иногда необходимо присваивать каждому из одинаковых значений средний ранг, в данном случае — 5,5. Например, такие ранги вычисляются в критерии Уилкоксона-Манна-Уитни для проверки гипотезы о равенстве математических ожиданий (см. раздел 2.4.2).

На рис. 6.11 показаны выборка и два метода ранжирования выборочных значений. В столбцах C, D, E ранги подсчитаны с помощью средства Ранг и перцентиль и затем отсортированы в порядке возрастания значений столбца Точка (столбец Процент, также генерируемый этим средством, удален). В столбце G для вычисления рангов используется формула массива, которая в ячейке G2 имеет следующий вид:

$\{=\text{ЕСЛИ}((\text{СУММ}(\text{ЕСЛИ}(\text{Выборка}=\text{A2};1)))=1;(\text{СУММ}(\text{ЕСЛИ}(\text{Выборка} \geq \text{A2};1;0)));(\text{СУММ}(\text{ЕСЛИ}(\text{Выборка} \geq \text{A2};1)))-((\text{СУММ}(\text{ЕСЛИ}(\text{Выборка}=\text{A2};1))-1)*0,5)\}$.

Эта формула введена в ячейку G2 как формула массива и затем скопирована в ячейки, расположенные ниже. Формула, на первый взгляд, кажется довольно сложной, но, разбив ее на отдельные части, в ней нетрудно разобраться.

На рисунке отмечены ранги одинаковых значений, вычисляемые средством Ранг и перцентиль и данной формулой массива.

В заключение отметим, что основное преимущество использования формул массивов по сравнению со средствами пакета анализа (типа Ранг и перцентиль) заклю-

чается в том, что эти формулы динамичны и сразу выдают значения при изменении входных данных. Кроме того, формулы массивов часто исключают необходимость использования промежуточных формул (см. пример из предыдущего раздела). И, в конце концов, они позволяют выполнять вычисления, которые трудно или невозможно сделать по-другому. Конечно, формулы массивов имеют и свои недостатки, среди которых отметим трудность их понимания и то, что формулы массивов нельзя экспортировать в форматы других программ электронных таблиц, например программы Lotus 1-2-3.

	A	B	C	D	E	F	G	H	I
1	Выборка		Точка	Выборка	Ранг		Формула массива		
2	93		1	93	3		3		
3	95		2	95	2		2		
4	30		3	30	14		14,5		
5	30		4	30	14		14,5		
6	97		5	97	1		1		
7	4		6	4	19		19		
8	75		7	75	7		7		
9	52		8	52	10		10		
10	58		9	58	9		9		
11	83		10	83	4		5		
12	83		11	83	4		5		
13	83		12	83	4		5		
14	64		13	64	8		8		
15	6		14	6	17		17		
16	5		15	5	18		18		
17	1		16	1	20		20		
18	17		17	17	16		16		
19	41		18	41	11		11,5		
20	41		19	41	11		11,5		
21	36		20	36	13		13		
22									

Рис. 6.11. Ранжирование данных с помощью формулы массива

Закончим раздел, посвященный массивам и формулам массивов, описанием функций Excel, позволяющих работать с матрицами (т.е. с теми же массивами), а также функций суммирования, в частности суммирования произведений значений двух массивов, которые часто используются при проведении статистического анализа для выполнения самых разнообразных вычислений.

6.1.5. Матричные вычисления

Функции для работы с матрицами следующие.

Функция	Назначение
МОБР	Возвращает обратную матрицу
МОПРЕД	Возвращает определитель матрицы
МУМНОЖ	Возвращает произведение матриц

Синтаксис функций:

ФУНКЦИЯ(Массив1;Массив2)

Функции МОБР и МОПРЕД имеют по одному аргументу Массив, а функция МУМНОЖ — два. Аргумент Массив может быть задан как диапазон ячеек, как массив констант или как имя диапазона или массива. Если какая-либо из ячеек в массиве пуста или содержит текст, то функции возвращают значение ошибки #ЗНАЧ!.

В функциях МОБР и МОПРЕД аргумент Массив должен иметь равное количество строк и столбцов, поскольку эти функции работают только с квадратными матрицами. Если Массив имеет неравное число строк и столбцов, то функции возвращают значение ошибки #ЗНАЧ!. Если определитель обращаемой матрицы равен нулю, то в этом случае функция МОБР возвращает значение ошибки #ЧИСЛО!.

Функция МУМНОЖ выполняет умножение матриц стандартным образом и требует двух аргументов: одной матрицы размером $n \times k$ (n — количество строк, k — количество столбцов) и второй матрицы размером $k \times m$ (здесь также k — количество строк, m — количество столбцов). Результирующая матрица будет иметь размер $n \times m$. Таким образом, количество столбцов аргумента Массив1 должно быть таким же, как количество строк аргумента Массив2. Если это условие не выполняется, то функция возвращает значение ошибки #ЗНАЧ!.

Функции МОБР и МУМНОЖ должны вводиться, как формулы массивов, т.е. с использованием комбинации клавиш <Ctrl+Shift+Enter> (предварительно следует выделить диапазон ячеек, в котором будет выведен результат вычислений).

Все три функции, которые здесь рассматриваются, производят вычисления с точностью до 16 значащих цифр, что может привести к небольшим численным ошибкам округления. Поэтому числа, значения которых имеют порядок $1\text{E}-16$ или меньше, можно считать нулевыми.

На рис. 6.12 показаны решение системы линейных алгебраических уравнений и формулы, по которым находится это решение. Такую схему вычислений можно использовать, например, для нахождения коэффициентов уравнения регрессии (см. раздел 3.4). Перемножение матрицы системы и обратной к ней матрицы сделано для того, чтобы оценить точность вычислений, поскольку результатом этого произведения матриц должна быть единичная матрица, внедиагональные элементы которой равны нулю.

6.1.6. Функции суммирования

В Excel имеется богатый арсенал функций суммирования, многие из которых можно использовать для вычисления статистических характеристик выборок. Мы не будем описывать "известные" функции суммирования СУММ и СУММЕСЛИ; рассмотрим другие, "менее известные", функции.

Функция СУММКВ

Функция вычисляет сумму квадратов своих аргументов, т.е. вычисляет сумму вида $\sum_i x_i^2$, где x_i — значения массива. Ее можно использовать, например, при вычислении выборочной дисперсии или сумм квадратов в дисперсионном анализе (см. раздел 3.5).

	A	B	C	D	E	F	G
1	Матрица коэффициентов системы				Вектор правых частей уравнений		Определитель матрицы системы
2	1	1	1	4	15		-216
3	-2	3	4	0	11		=МОПРЕД(A2:D5)
4	0	2	-5	1	-5		
5	3	-5	0	7	16		
6							
7	Обратная матрица				Решение		Точное решение
8	0.837963	-0.713	-0.4028	-0.4213	-3.55271E-15		0
9	0.3796296	-0.1296	-0.0278	-0.21296			1
10	0.1342593	-0.0093	-0.1806	-0.05093			2
11	-0.087963	0.21296	0.15278	0.171296			3
12	Формулы: {=МОБР(A2:D5)}				{=МУМНОЖ(A8:D11;E2:E5)}		
13	Произведение матрицы системы						
14	и обратной к ней матрицы:						
15	1	-2E-16	0	0	{=МУМНОЖ(A2:D5;A8:D11)}		
16	-3.33E-16	1	1.1E-16	3.05E-16			
17	-9.71E-17	-3E-17	1	8.33E-17			
18	-1.1E-16	0	0	0			
19							

Рис. 6.12. Решение системы линейных алгебраических уравнений

Синтаксис функции:

СУММКВ(Число1;Число2;...)

Может иметь до 30 аргументов Число. В качестве аргументов можно использовать массивы и ссылки на массивы. Если среди аргументов имеются текстовые или логические значения, то они игнорируются.

Функция СУММКВРАЗН

Функция вычисляет сумму квадратов попарных разностей значений двух массивов, т.е. вычисляет сумму вида $\sum_i (x_i - y_i)^2$, где x_i и y_i — значения массивов.

Синтаксис функции:

СУММКВРАЗН(Массив_х,-Массив_у)

Аргументы Массив — числа, массивы или ссылки на диапазоны ячеек. Если аргумент Массив содержит текстовые или логические значения либо пустые ячейки, то такие значения игнорируются; однако ячейки, которые содержат нулевые значения, учитываются.

Если аргументы Массив_х и Массив_у содержат различные количества элементов, то функция возвращает значение ошибки #Н/Д.

Функция СУММПРОИЗВ

Функция перемножает соответствующие элементы заданных массивов и возвращает сумму этих произведений, т.е. вычисляет сумму вида $\sum_i x_i y_i z_i$, где x_i , y_i и z_i — значения массивов.

Синтаксис функции:

СУММПРОИЗВ(Массив1;Массив2;Массив3;...)

Функция может иметь до 30 аргументов. Аргументы Массив должны иметь одну и ту же размерность, т.е., если хотя бы один аргумент является отдельным числом, то все остальные аргументы должны быть числами. Если же аргумент Массив1 является массивом или ссылкой на диапазон ячеек, все остальные аргументы должны иметь такую же размерность, что и аргумент Массив1. В противном случае функция возвращает значение ошибки #ЗНАЧ!. Нечисловые элементы аргументов трактуются как нулевые. Если задан только один аргумент, возвращается сумма элементов этого аргумента.

Функция СУММРАЗНKB

Функция возвращает сумму попарных разностей квадратов соответствующих значений двух массивов, т.е. вычисляет сумму вида $\sum_i (x_i^2 - y_i^2)$, где x_i и y_i — значения массивов.

Синтаксис функции:

СУММРАЗНKB(Массив_х;Массив_у)

Аргументы функции должны быть числами, массивами или ссылками на диапазоны ячеек, содержащие числа. Текстовые и логические значения, а также пустые ячейки в массивах и диапазонах игнорируются. Если аргументы Массив_х и Массив_у имеют различные количества элементов, функция возвращает значение ошибки #Н/Д.

Функция СУММСУММКВ

Возвращает сумму попарных сумм квадратов соответствующих элементов двух массивов, т.е. вычисляет сумму вида $\sum_i (x_i^2 + y_i^2)$, где x_i и y_i — значения массивов.

Синтаксис функции:

СУММСУММКВ(Массив_х;Массив_у)

Аргументы функции должны быть числами, массивами или ссылками на диапазоны ячеек, содержащие числа. Текстовые и логические значения, а также пустые ячейки в массивах и диапазонах игнорируются. Если аргументы Массив_х и Массив_у имеют различные количества элементов, функция возвращает значение ошибки #Н/Д.

Различие между одинаковыми формулами =СУММСУММКВ(Х;У) и =СУММКВ(Х) + СУММКВ(У) проявляется только тогда, когда в массиве Х или У имеются элементы, которые игнорируются функциями (т.е. текстовые, логические или пустые ячейки). В первой формуле из суммы исключается пара элементов, принадлежащих массивам Х и У, если хотя бы один из них игнорируется. Во второй формуле подобные исключения из сумм происходят независимо.

6.2. Диаграммы

Предполагая, что читатель знаком с основами построения и применения диаграмм и графиков в Excel, здесь мы рассмотрим только некоторые их возможности, а именно — добавление линий тренда, добавление планок погрешностей и создание гистограмм распределений.

6.2.1. Линии тренда

Добавляя линию тренда к построенному графику изменения данных, можно оценить динамику изменения этих данных. Линия тренда — это уравнение регрессии (см. раздел 3.4), которое строится методом наименьших квадратов на основании существующих рядов данных и может быть экстраполировано за интервал исходных данных.

Линии тренда могут быть добавлены только к определенным типам диаграмм: к диаграммам с областями, гистограммам, графикам, линейчатым и точечным диаграммам.

Для добавления к диаграмме линии тренда выполните следующие действия.

1. Выделите диаграмму или ряд данных, к которым необходимо добавить линию тренда.
2. Из меню Диаграмма выберите команду Добавить линию тренда, и откроется диалоговое окно Линия тренда (рис. 6.13).
3. На вкладке Тип выберите подходящий тип линии тренда.
4. Щелкните на кнопке ОК.

Пример квадратичной линии тренда показан на рис. 6.14. Обращаем внимание, что дополнительно можно вывести уравнение регрессии и коэффициент детерминации D^2 .

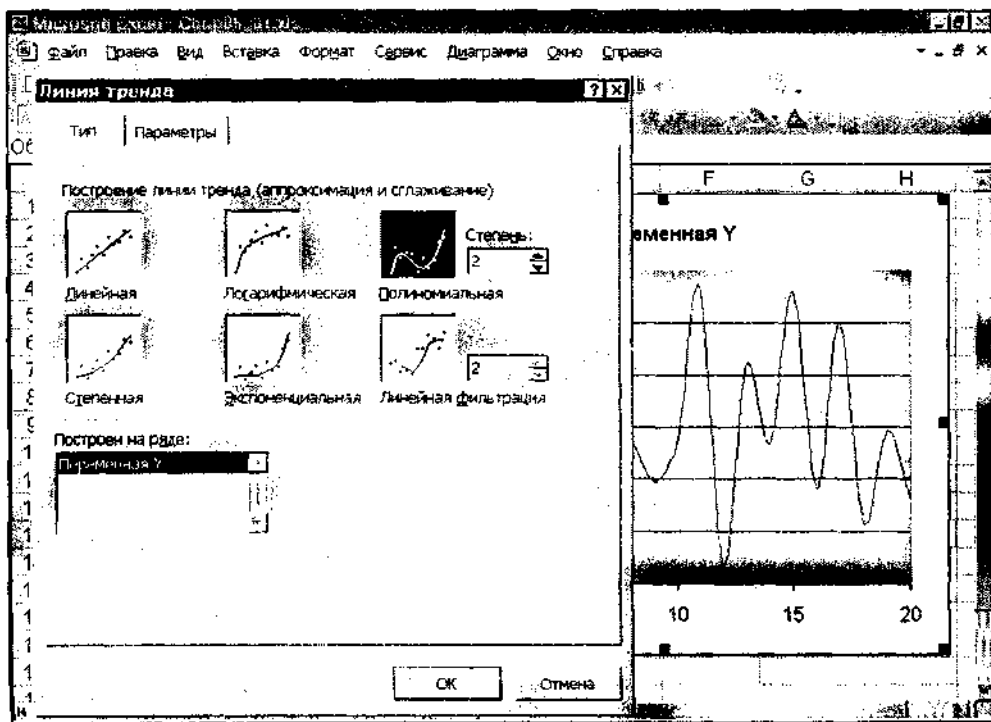


Рис. 6.13. Диалоговое окно Линия тренда

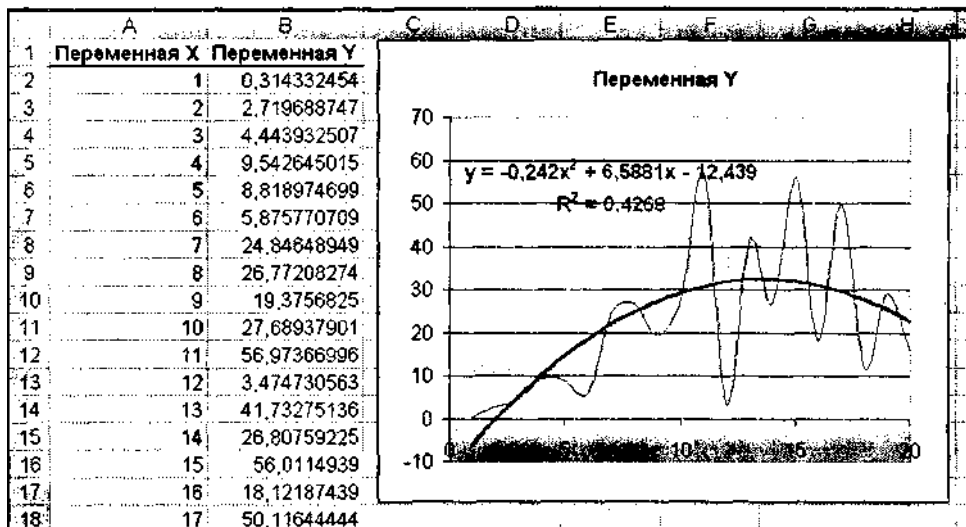


Рис. 6.14. Пример линии тренда

На вкладке Тип диалогового окна Линия тренда можно выбрать следующие типы линий тренда (более подробная информация об этих типах регрессии приведена в разделах 3.4 и 15.1).

- Линейная. Строит прямую линию на основании рассчитанного уравнения линейной регрессии.
- Логарифмическая. Строит логарифмическую линию на основании рассчитанного уравнения нелинейной регрессии, использующей логарифмический тип зависимости.
- Полиномиальная. Строит полиномиальную линию на основании рассчитанного уравнения нелинейной регрессии, использующей полиномиальный тип зависимости. В поле Степень задается степень аппроксимирующего полинома (значение может лежать в интервале от 2 до 6).
- Степенная. Строит степенную линию на основании рассчитанного уравнения нелинейной регрессии, использующей степенной тип зависимости.
- Экспоненциальная. Строит экспоненциальную линию на основании рассчитанного уравнения нелинейной регрессии, использующей экспоненциальный тип зависимости.
- Линейная фильтрация. Строит кривую методом скользящего среднего. Количество точек на этой кривой равно числу точек ряда данных минус число, указанное в поле Точки. В этом поле задается количество точек, используемых для вычисления скользящего среднего. (О методе скользящего среднего речь идет в разделе 5.17.)

На этой же вкладке имеется список Построен на ряде, в котором указывается, на основе какого ряда данных строится линия тренда.

Параметры линии тренда

Опции, приведенные на вкладке Параметры диалогового окна Линия тренда (рис. 6.15), предлагают следующие возможности.

- Название аппроксимирующей (сглаженной) кривой. Опции этой группы позволяют задать название линии тренда, либо выбрав предлагаемое по умолчанию (переключатель автоматическое), либо введя свой вариант имени (переключатель другое).
- Прогноз. Опции этой группы задают число периодов (как назад, так и вперед), на которые будет выполнена экстраполяция линии тренда.
- Опция пересечение кривой с осью Y в точке определяет точку пересечения линии тренда с осью Y.
- Опции показывать уравнение на диаграмме и поместить на диаграмму величину достоверности аппроксимации (R^2) используются для отображения на диаграмме уравнения регрессии и значения коэффициента детерминации, как показано на рис. 6.14.

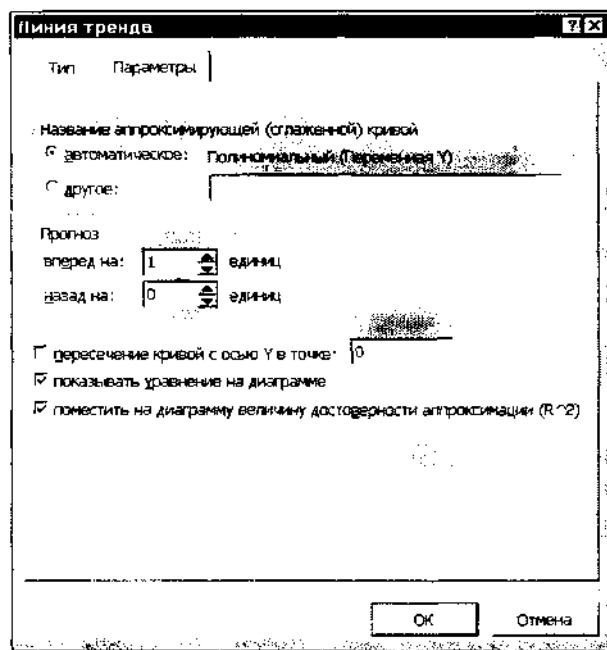


Рис. 6.15. Вкладка Параметры диалогового окна Линия тренда

Форматирование линии тренда

Как и в случае использования других элементов диаграммы, может возникнуть необходимость изменить какие-либо параметры отображения линии тренда. Выделив линию тренда, можно форматировать ее точно так же, как и любой другой элемент диаграммы.

Для форматирования линии тренда выполните следующее.

1. Выделите линию тренда.
2. Из меню **Формат** выберите команду **Выделенная линия тренда**, и откроется диалоговое окно **Формат линии тренда**.
3. На вкладке **Вид** диалогового окна **Формат линии тренда** выберите необходимые параметры форматирования (рис. 6.16).

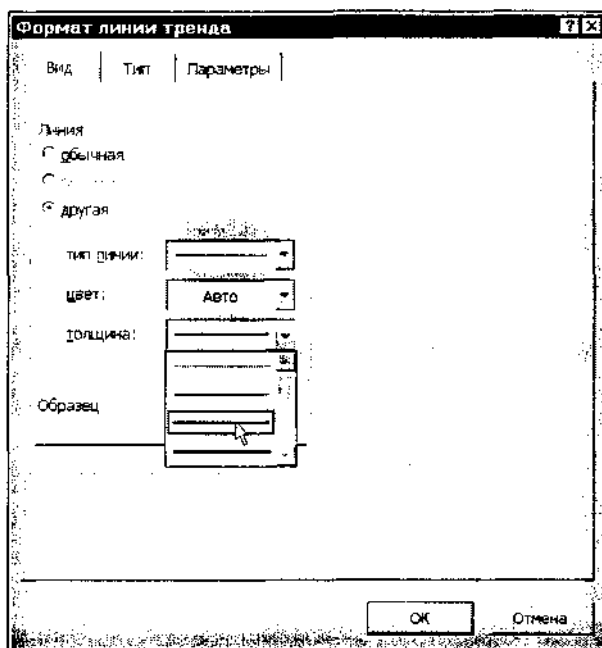


Рис. 6.16. Вкладка Вид диалогового окна Формат линии тренда

Две другие вкладки диалогового окна **Формат линии тренда** повторяют вкладки **Тип** и **Параметры** диалогового окна **Линия тренда**. Поэтому в окне **Формат линии тренда** можно также изменить тип линии тренда и ее параметры.

6.2.2. Планки погрешностей

Для определенных типов диаграмм можно добавить к точкам данных планки погрешностей. Они обычно используются для того, чтобы показать степень изменчивости значения данных в конкретной точке. Планки погрешностей для значений оси **Y** применяются только для диаграмм с областями, гистограмм, графиков, линейчатых и точечных диаграмм. Набор данных точечной диаграммы может иметь планки погрешностей и для значений оси **X**, и для значений оси **Y** одновременно. Щелкните в диалоговом окне **Формат ряда данных** на вкладке **Y-погрешности**, чтобы вывести на экран опции, показанные на рис. 6.17.

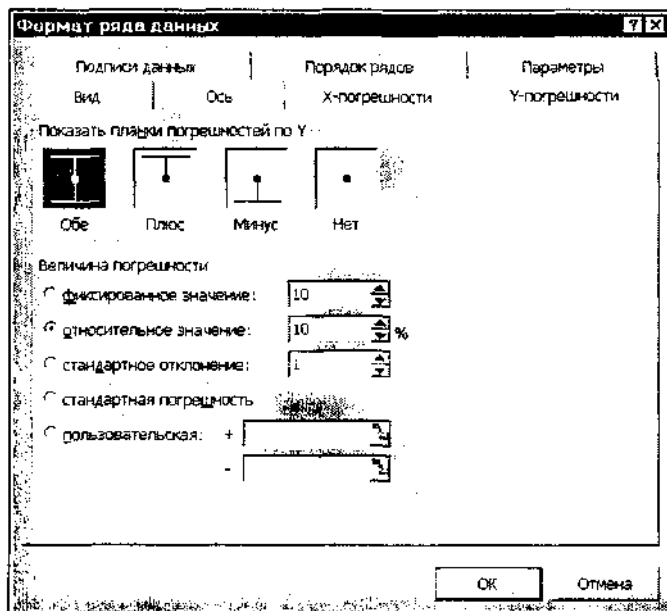


Рис. 6.17. Вкладка Y-погрешности диалогового окна Формат ряда данных

В Excel можно установить следующие типы планок погрешностей.

- Переключатель фиксированное значение. Планки погрешностей имеют заданный фиксированный размер.
- Переключатель относительное значение. Размер планки погрешностей задается в процентах от каждого значения.
- Переключатель стандартное отклонение. Размер планки погрешностей задается в единицах среднеквадратического отклонения от среднего значения, которые Excel вычисляет для ряда данных.
- Переключатель стандартная погрешность. Размер планки погрешностей задается в единицах среднеквадратической ошибки, которую вычисляет Excel для ряда данных.
- Переключатель пользовательская. Здесь можно указать размер верхней и нижней планок погрешностей. В поля ввода этой опции можно ввести значения или ссылку на диапазон, в котором содержатся значения погрешностей.

На рис. 6.18 показана диаграмма, к которой были добавлены планки погрешностей. Можно открыть диалоговое окно Формат планок погрешностей и с его помощью изменить внешний вид планок погрешностей, например стиль линий и цвет.

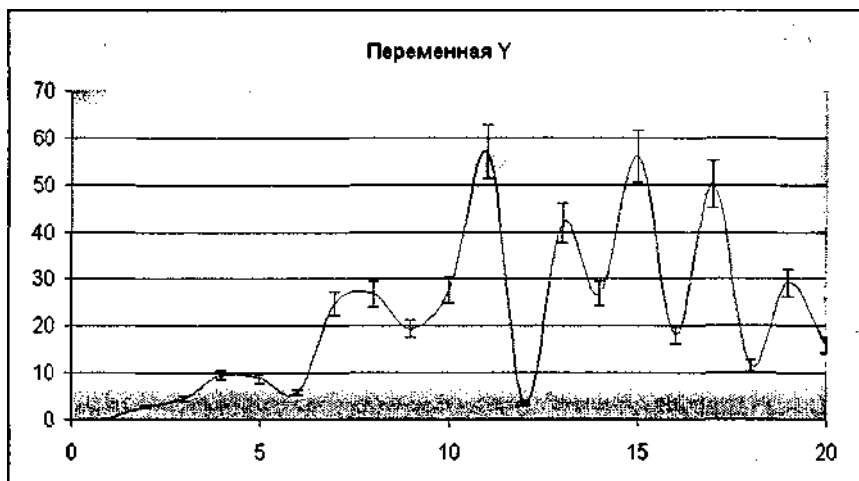


Рис. 6.18. На диаграмме к ряду данных добавлены планки погрешностей

6.2.3. Построение гистограмм и функций распределения дискретных случайных величин

Обычно для построения гистограмм дискретных распределений в Excel используется тип диаграммы Гистограмма. Однако этот тип диаграммы можно использовать только тогда, когда значения, принимаемые дискретной случайной величиной, располагаются на оси ОХ равномерно (т.е. через равные промежутки). Если же они располагаются неравномерно, то подходит только тип диаграммы Точечная. Но этот тип диаграммы не позволяет строить столбцовые диаграммы. Отметим, что средство Гистограмма из пакета анализа (см. раздел 5.2) также не может строить гистограммы для неравномерно распределенных значений.

Из этой ситуации предлагаем следующий выход. Пусть имеется вероятностная таблица, в столбце А содержащая значения, которые может принимать случайная величина, а в столбце В — вероятности принятия этих значений, как показано на рис. 6.19. Строится диаграмма типа Точечная без линий, соединяющих точки данных. Затем выделяется ряд данных и выбирается команда Формат^Выделенный ряд. В открывшемся диалоговом окне Формат ряда данных

	А	В	С
1	Значения	Вероятность	
2	1,71	0,24	
3	1,98	0,09	
4	2,48	0,2	
5	4,1	0,08	
6	4,44	0,06	
7	4,5	0,09	
8	4,56	0,24	
9			
10			

Рис. 6.19. Частотная таблица

переходим на вкладку Y-погрешности и задаем планку погрешности типа Минус. В качестве величины погрешности задаем Относительное значение 100% (рис. 6.20). На графике появляются вертикальные столбцы от значений данных до оси Х. Теперь остается отформатировать планки погрешностей и значения данных. У значений данных убираем маркеры (опция Отсутствует в области Маркер вкладки Вид диалогового окна Формат ряда данных), для планок погрешностей в диалоговом окне Формат планок погрешностей на вкладке Вид выбираем вид

планки без горизонтальной линии и делаем ее максимально "толстой" (эти опции показаны на рис. 6.21). В результате получаем гистограмму выборки, показанную на рис. 6.22.

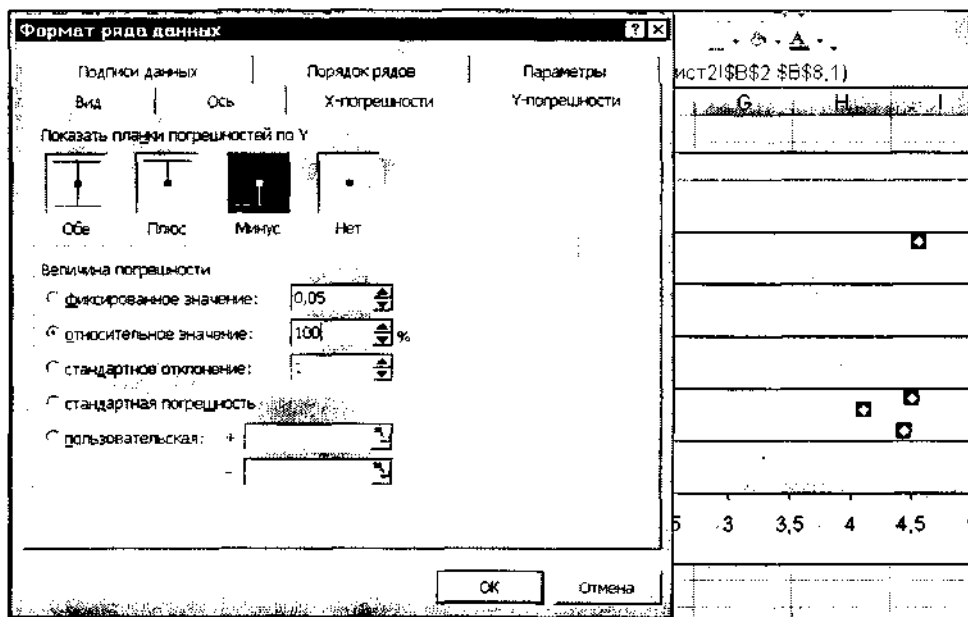


Рис. 6.20. Диалоговое окно Формат ряда данных

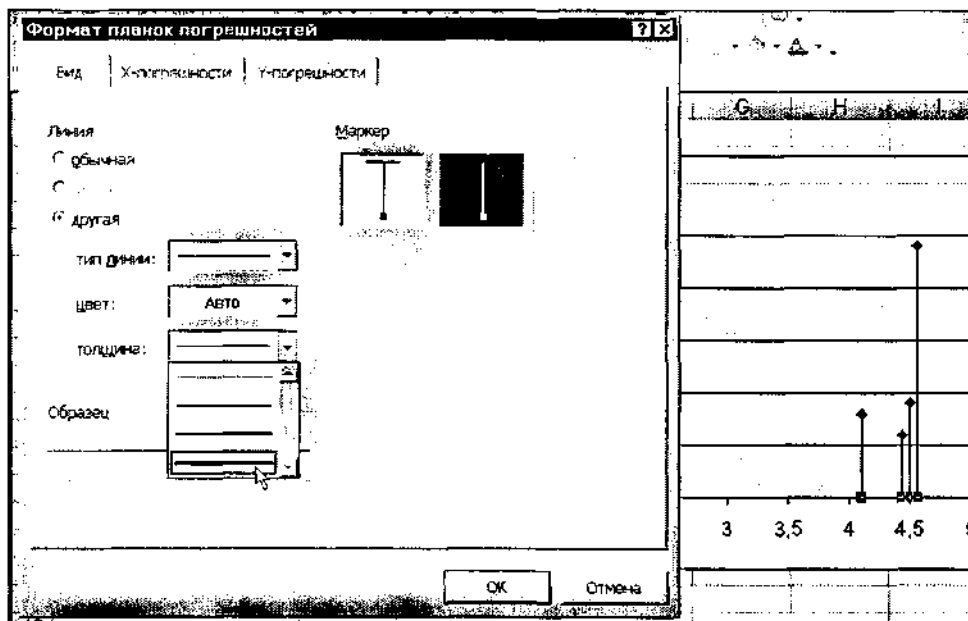


Рис. 6.21. Диалоговое окно Формат планок погрешностей

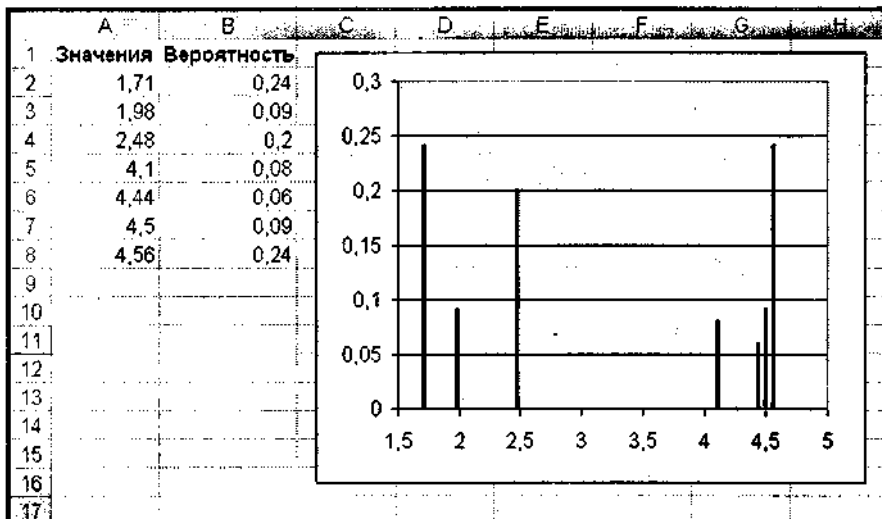


Рис. 6.22. Гистограмма для дискретного распределения

Такую диаграмму можно сохранить как нестандартную в средстве построения диаграмм Excel для дальнейшего использования. Для этого выделите диаграмму, выполните команду **Диаграмма^Тип диаграммы**, в открывшемся диалоговом окне **Тип диаграммы** перейдите на вкладку **Нестандартные**, установите переключатель **Дополнительные** и щелкните на кнопке **Добавить**. В открывшемся диалоговом окне **Добавление нового типа диаграммы** присвойте имя этому типу диаграммы и дайте ее описание.

Построение эмпирической функции распределения для дискретных случайных величин также имеет некоторые сложности, поскольку такая функция имеет ступенчатый вид, но ни средство построения диаграмм Excel, ни средство **Гистограмма** из пакета анализа такие графики строить не может. Покажем, как все-таки в Excel построить такой график.

Пусть заданы значения функции распределения, как показано на рис. 6.23. Чтобы подсчитать эти значения на основе известных вероятностей отдельных значений, в ячейку C2 (если данные располагаются так же, как на рис. 6.23) записывается формула $=B2$, в ячейку C3 — $=C2+B3$. Затем эта формула распространяется вниз до ячейки C8.

Чтобы построить ступенчатый график функции распределения, некоторые операции придется выполнить вручную. Сначала вставим пустой столбец перед столбцом, содержащим значения функции распределения, и скопируем в него выборочные значения из столбца A. Затем перед каждой строкой в столбцах C и D (теперь в столбце D находятся значения функции распределения) вставим по пустой строке, сдвигая ячейки вниз. Должно получиться так, как показано на рис. 6.24.

Далее в ячейку C2 введем формулу $=C3-0,0000001$, а в ячейку D2 — число 0. Формулу из ячейки C2 скопируем в ячейку C4, а в ячейку D4 введем формулу $=D3$. Выделим ячейки C4:D4 и скопируем их во все свободные ячейки вниз до строки 14. В ячейку C16 можно ввести число 5, а в ячейку D16 — **число 1** (но это не обязательно). Рабочий лист на этом этапе показан на рис. 6.25.

	C8		=C7+B8
	A	B	C
1	Значения	Вероятности	Функция распределения
2	1,71	0,24	0,24
3	1,98	0,09	0,33
4	2,48	0,2	0,53
5	4,1	0,08	0,61
6	4,44	0,06	0,67
7	4,5	0,09	0,76
8	4,56	0,24	1
9			

Рис. 6.23. Вычисление значений функции распределения

	A	B	C	D
	Значения	Вероятности	Значения	Функция распределения
1				
2	1,71	0,24		
3	1,98	0,09	1,71	0,24
4	2,48	0,2		
5	4,1	0,08	1,98	0,33
6	4,44	0,06		
7	4,5	0,09	2,48	0,53
8	4,56	0,24		
9			4,1	0,61
10				
11			4,44	0,67
12				
13			4,5	0,76
14				
15			4,56	1
16				

Рис. 6.24. Вставка пустых ячеек

Теперь для построения графика эмпирической функции распределения достаточно построить средствами Excel диаграмму типа Точечная с соединительными линиями без маркеров на основе данных диапазона C2:D16. Готовая отформатированная диаграмма показана на рис. 6.26. Чтобы провести пунктирные линии в узловых точках графика, используются планки погрешностей, как описано выше, при построении гистограммы. Такую диаграмму можно сохранить как нестандартную для дальнейшего использования.

6.2.4. Гистограммы с перекрытием

То, что будет показано в этом разделе, относится к "маленьким секрета" форматирования диаграмм и не играет принципиальной роли, однако позволяет сделать, например, гистограммы частот более наглядными для сравнения. Обычно на гистограммах, построенных по нескольким рядам данных, столбцы, соответствующие разным рядам данных, имеют одну и ту же ширину, определяемую параметром Ширина зазора на вкладке Параметры диалогового окна Формат ряда

данных. Если совместить столбцы разных рядов данных, то они перекрывают друг друга и диаграмма становится нечитаемой. Чтобы сделать столбцы разной ширины, как показано на рис. 6.27, значения одного или нескольких рядов данных следует отложить на *дополнительной оси*, а затем сделать эту ось невидимой и отменить заливку для одного из рядов данных.

	A	B	C	D
	Функция			
1	Значения	Вероятности	Значения	распределения
2	1,71	0,24	1,7099999	0
3	1,98	0,09	1,71	0,24
4	2,48	0,2	1,9799999	0,24
5	4,1	0,08	1,98	0,33
6	4,44	0,06	2,4799999	0,33
7	4,5	0,09	2,48	0,53
8	4,56	0,24	4,0999999	0,53
9			4,1	0,61
10			4,4399999	0,61
11			4,44	0,67
12			4,4999999	0,67
13			4,5	0,76
14			4,5599999	0,76
15			4,56	1
16			5	1
17				

Рис. 6.25. Все готово для построения графика

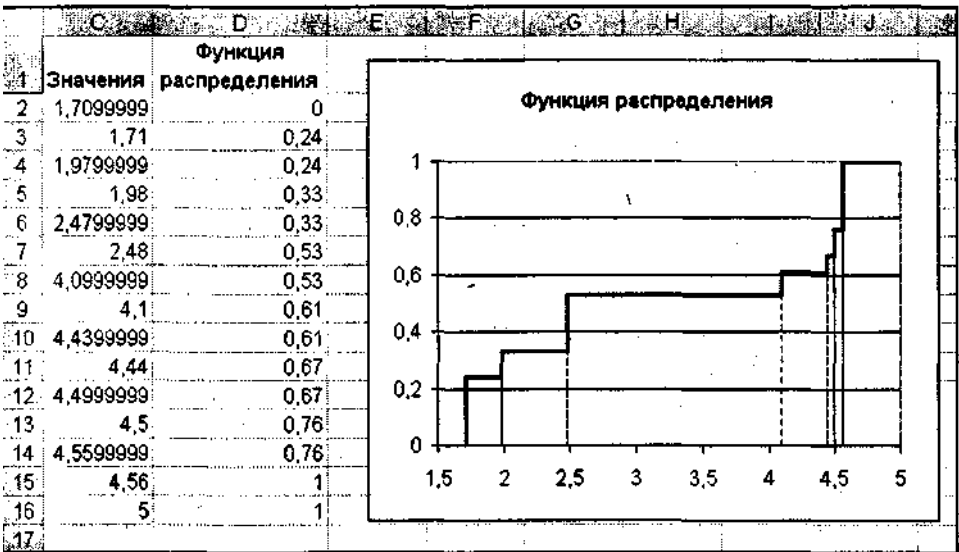


Рис. 6.26. Эмпирическая функция распределения

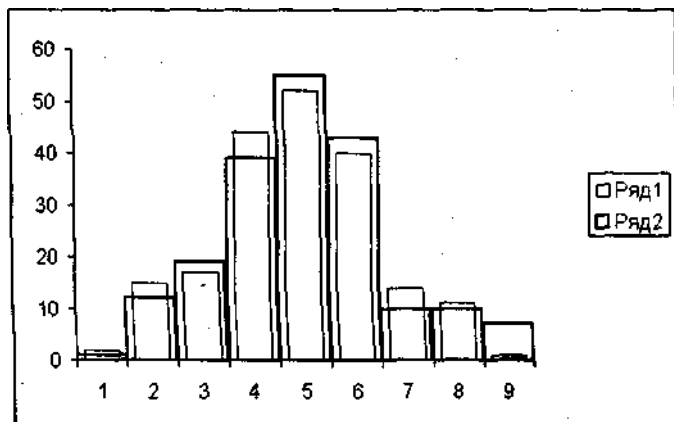


Рис. 6.27. Гистограмма с перекрытием

Для создания подобного эффекта выполните следующее.

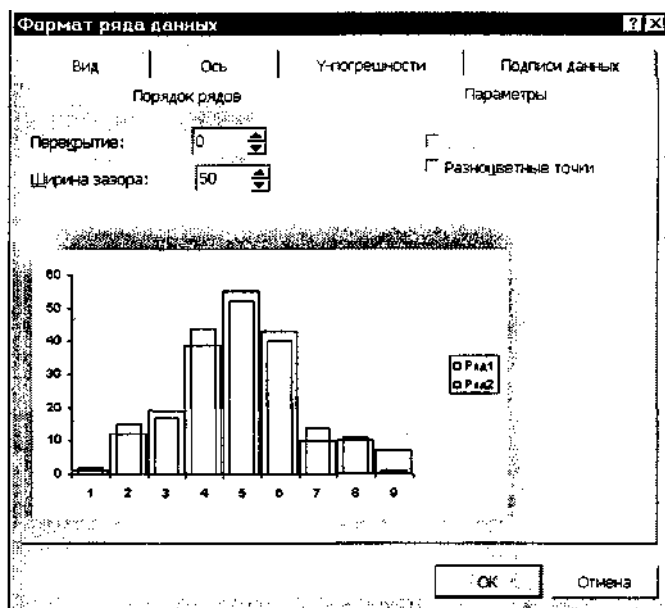
1. Начните с построения гистограммы.
2. Выделите один из рядов данных.
3. Из меню **Формат** выберите команду **Выделенный ряд**, и откроется диалоговое окно **Формат ряда данных**.
4. Перейдите на вкладку **Вид** и в группе **Заливка** установите переключатель **прозрачная**.
5. Перейдите на вкладку **Ось** и установите переключатель по вспомогательной оси.
6. Перейдите на вкладку **Параметры** (рис. 6.28).
7. Присвойте параметру **Перекрытие** значение 0, а параметру **Ширина зазора** — значение 50.
8. Щелкните на кнопке **ОК**.

6.3. Настройка Поиск решения

Поиск решения — это надстройка, входящая в поставку Excel и предназначенная для решения задач линейной и нелинейной оптимизации¹. Для этого в ней используются методы и алгоритмы математического программирования, которые позволяют находить оптимальные решения задач оптимизации, представленных в Excel в виде табличных моделей. Для линейных задач надстройка Поиск решения использует симплекс-метод, для задач целочисленного программирования —

¹ *Надстройка Поиск решения (в оригинальной англоязычной версии Excel она называется Solver) разработана компанией Frontline Systems (<http://www.frontsys.com>). Этой же компанией разработано несколько расширенных коммерческих программ-оптимизаторов, в том числе надстройка Premium Edition Solver, которая не только расширяет возможности стандартной надстройки Поиск решения (например, содержит несколько способов коррекции ошибок и генерирует дополнительные отчеты), но и имеет несколько новых встроенных алгоритмов решения существенно нелинейных задач, в том числе генетический алгоритм.*

метод ветвей и границ и для нелинейных задач — метод приведенного градиента. Подробные сведения о настройке Поиск решения и ее использовании для решения оптимизационных задач, можно найти в книге [12].



Рмс. 6.2S. Установка значения перекрывания и ширины зазора

Средство Поиск решения можно эффективно использовать не только при решении задач оптимизации, но и при проведении статистического анализа. Ниже будут показаны примеры применения средства Поиск решения для решения систем линейных алгебраических уравнений (можно использовать для вычисления коэффициентов уравнений регрессии) и для подбора параметров распределений. Другие применения средства Поиск решения для моделирования случайных величин показаны в главе 7.

6.3.1. Задачи оптимизации и средство Поиск решения

Поскольку рассматриваемое здесь средство предназначено, в первую очередь, для решения задач оптимизации, необходимо иметь хотя бы общее представление об этих задачах и знать соответствующую терминологию, так как она используется при задании параметров данного средства. Поэтому приведем общую формулировку задач оптимизации и покажем, как представить ее в виде табличной модели на рабочем листе Excel.

Общую задачу оптимизации можно сформулировать следующим образом. Пусть $\mathbf{X} = (x_1, x_2, \dots, x_n)$ — вектор действительных переменных. Необходимо

минимизировать (или максимизировать) целевую функцию $z = f(\mathbf{X})$

при выполнении ограничений

$$g_1(\mathbf{X}) \leq b_1,$$

$$g_2(X) \leq b_2,$$

...

$$g_m(X) \leq b_m.$$

Обычно предполагается, что функции $f(X)$ и $g_i(X)$ ($i = 1, 2, \dots, m$) дважды непрерывно дифференцируемы. Часто добавляются условия неотрицательности переменных $X > 0$, которые могут как включаться в указанные m ограничений, так и не включаться. Среди ограничений могут быть ограничения в виде неравенств и в виде равенств. Вектор $\{b_1, b_2, \dots, b_m\}$ называется вектором правых частей ограничений.

Если все функции $f(X)$ и $g_i(X)$ линейны относительно переменных x_1, x_2, \dots, x_n , Имеем задачу линейной оптимизации; если хотя бы одна из этих функций нелинейная, получаем задачу нелинейной оптимизации.

Итак, задача оптимизации включает три "объекта": переменные x_1, x_2, \dots, x_n (в средстве Поиск решения ячейки, содержащей значения этих переменных, они называются *изменяемыми ячейками*), целевая функция (ячейка, содержащая значение этой функции в средстве Поиск решения, называется *целевой ячейкой*) и ограничения (для применения средства Поиск решения ограничения могут быть записаны на рабочем листе и затем указаны в диалоговом окне этого средства либо заданы непосредственно в этом окне без записи на рабочем листе). При задании ограничений отдельно указываются функции ограничений $g_i(X)$ ($i = 1, 2, \dots, m$) и вектор правых частей ограничений (b_1, b_2, \dots, b_m) .

При создании табличной модели оптимизации в Excel предлагаем учитывать следующие рекомендации, которые облегчат дальнейшее применение средства Поиск решения.

- Значения переменных располагаются в отдельных ячейках и группируются в отдельный блок ячеек.
- Каждому ограничению отводится отдельная строка или столбец таблицы. Ограничения группируются в отдельный блок ячеек.
- Желательно, чтобы ячейки, содержащие переменные и значение целевой функции, а также все ограничения, имели заголовки.
- Коэффициенты целевой функции должны храниться в отдельной строке, располагаясь непосредственно под или над соответствующими переменными; формула для вычисления целевой функции должна находиться в соседней ячейке.
- В каждой строке ограничений за ячейками, содержащими коэффициенты данного ограничения, следует ячейка, в которую записывается вычисленное значение функции ограничения (значение левой части ограничения). За ней может следовать ячейка, в которой стоит соответствующий знак неравенства или равенства ограничения, а затем ячейка, содержащая значение правой части ограничения. Дополнительно можно иметь ячейку, в которой вычислена разность между значениями левой и правой частей неравенства.
- Условия неотрицательности переменных решения не обязательно включать в табличную модель. Как правило, они опускаются и указываются непосредственно в диалоговом окне средства Поиск решения.

В результате выполнения этих рекомендаций все основные коэффициенты модели содержатся в отдельных ячейках, поэтому их легко изменять, не меняя формул модели. Благодаря группированию упрощается работа со средством Поиск решения, поскольку для указания переменных или ограничений можно использовать *диапазоны ячеек*, т.е. задавать переменные и ограничения группой, а не по отдельности.

На рис. 6.29 показана табличная модель для следующей простой задачи:

$$\text{минимизировать } z = 2x_1 + 3x_2 + 5x_3$$

при ограничениях

$$\begin{aligned} x_1 + x_2 - x_3 &\geq -5, \\ -6x_1 + 7x_2 - 9x_3 &\leq 4, \\ x_1 + x_2 + 4x_3 &= 10. \end{aligned}$$

На переменные x_1 и x_2 также накладываются условия неотрицательности².

	A	B	C	D	E	F	G	H	I
1	Линейная задача оптимизации								
2		x1	x2	x3					
3	Переменные	1	2	3	Значение целевой функции				
4	Коэффициенты целевой функции	2	3	5	23	←=СУММПРОИЗВ(B3:D3;B4:D4)			
5	Ограничения	Коэффициенты функций ограничений			Значения функций ограничений	Правая часть ограничений		Разности	
6	Ограничение 1	1	1	-1	0	>=	-5	-5	
7	Ограничение 2	-6	7	-9	-18	<=	4	23	
8	Ограничение 3	1	1	4	15	=	10	-5	
9									
10		=СУММПРОИЗВ(\$B\$3:\$D\$3;E8:D8)						=G8-E8	
11									

Рис. 6.29. Табличная модель задачи оптимизации

На этом же примере нанесем первый "визит" к средству Поиск решения. Но сначала сделаем следующее замечание. Надстройка Поиск решения, хотя и входит в поставку Excel, не подключается автоматически к этой программе. Поэтому, если в меню Сервис нет команды Поиск решения, значит, надстройка не подключена. Для ее подключения надо выполнить команду Сервисе Надстройки и в открывшемся диалоговом окне Надстройки установить флажок перед опцией Поиск решения.

Надстройка Поиск решения используется следующим образом.

1. Откройте Excel и создайте табличную модель.
2. После отладки модели перейдите к этапу оптимизации, выбрав команду Поиск решения в меню Сервис.
3. В открывшемся диалоговом окне Поиск решения укажите данные, необходимые для процесса оптимизации (рис. 6.30).
 - а) В поле Установить целевую ячейку вводится адрес ячейки, содержащей значение целевой функции. Для нашей модели в это поле следует вве-

Здесь формулировка задачи преднамеренно не приведена к "правильному" виду (когда ограничения имеют один тип и т.п.), чтобы сделать этот маленький пример максимально обобщенным.

- сти E4, но лучше щелкнуть указателем мыши на этой ячейке, чтобы ввести ее адрес автоматически.
- б) Опции области Равной диалогового окна Поиск решения позволяют задать тип оптимизации. В данном случае необходимо минимизировать значение целевой функции. Для этого нужно щелкнуть на переключателе минимальному значению. Щелчок на переключателе максимальному значению укажет, что следует максимизировать целевую функцию. Можно также сделать значение целевой функции равным заданному числу, установив переключатель значению и введя это число.
 - в) Поле Изменяя ячейки позволяет указать ячейки, в которых содержатся переменные модели; в данном случае это диапазон В3:D3. (Можно попробовать воспользоваться кнопкой Предположить, но при этом обычно предлагаются неверные адреса ячеек переменных.)
 - г) Далее необходимо задать ограничения. Щелчок на кнопке Добавить открывает диалоговое окно Добавление ограничения, показанное на рис. 6.31. По умолчанию предполагается, что ограничение имеет вид неравенства со знаком <. Если табличная модель организована так, что неравенства одного знака расположены рядом, то их можно ввести все вместе, используя диапазоны ячеек. В противном случае придется вводить ограничения по отдельности, щелкая на кнопке Добавить диалогового окна Добавление ограничения. Заметим, что в поле Ссылка на ячейку нельзя вводить формулы — это должны быть ссылки на ячейки, которые, в свою очередь, могут содержать формулы.
 - д) Следует помнить об условиях неотрицательности для содержимого ячеек В3 и С3. Чтобы ввести эти ограничения, в диалоговом окне Добавление ограничения укажите диапазон В3:С3, выберите знак неравенства \geq и в поле Ограничение введите 0. Если условие неотрицательности накладывается на все переменные, то это условие можно задать в диалоговом окне Параметры поиска решения (опция Неотрицательные значения), которое открывается после щелчка на кнопке Параметры диалогового окна Поиск решения.
4. Поскольку мы работаем с линейной моделью, в диалоговом окне Параметры поиска решения установите флажок опции Линейная модель, а также Автоматическое масштабирование (рис. 6.32). Режим Автоматическое масштабирование предназначен для масштабирования числовых значений в модели таким образом, чтобы разность между наибольшим и наименьшим числами в модели была как можно меньшей, иначе в процессе вычислений могут возникнуть большие ошибки округления и результат может быть далеким от истинного решения. Остальные опции этого окна можно оставить без изменения — они, в основном, относятся к оптимизации целочисленных и нелинейных моделей. Щелкните на кнопке ОК, чтобы вернуться в диалоговое окно Поиск решения.
- б. После задания необходимых данных (указания ячейки, содержащей формулу для вычисления целевой функции, указания ячеек, в которых находятся переменные, и задания ограничений) щелкните на кнопке Выполнить.

6. Средство Поиск решения выполняет оптимизацию. В процессе вычислений в строке состояния отображаются число итераций и значения целевой функции при переборе множества допустимых решений задачи. Эта информация позволяет следить, как продвигается процесс оптимизации больших моделей, где он может длиться достаточно долго.
7. Если в табличной модели нет ошибок, Поиск решения выведет на экран диалоговое окно Результаты поиска решения (рис. 6.33), в котором можно указать, обновить ли исходную модель (т.е. занести ли в ячейки значения оптимального решения) и создавать ли отчет.

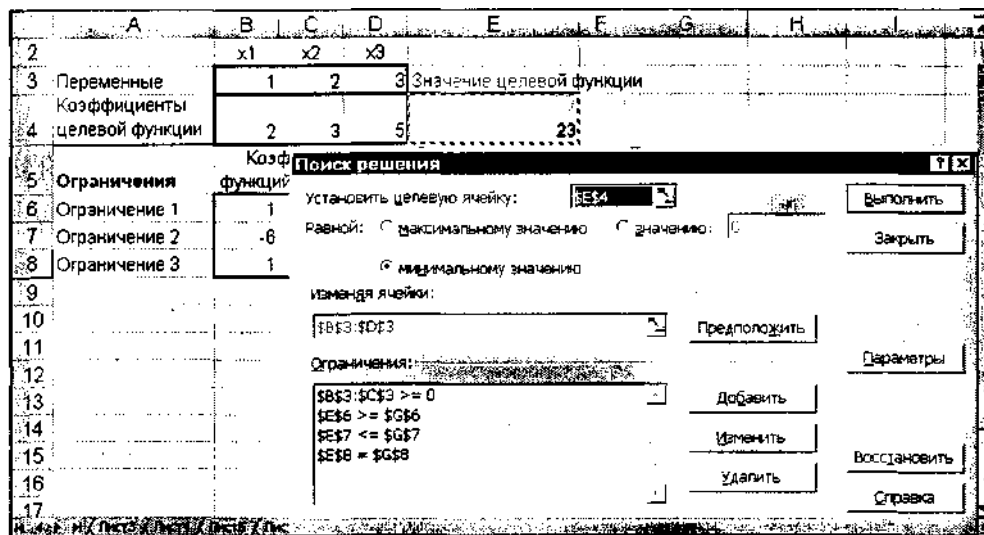


Рис. 6.30. Задание параметров для поиска решения

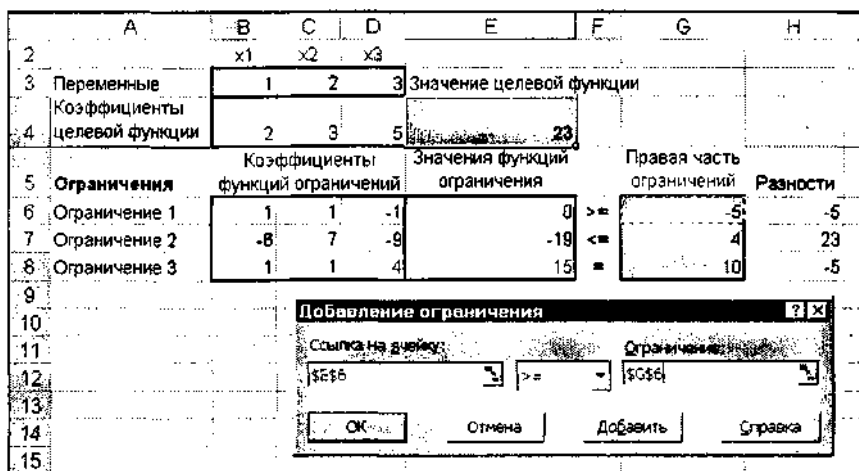


Рис. 6.31. Задание ограничений

	A	B	C	D	E	F	G	H	I
2		x1	x2	Параметры поиска решения					
3	Переменные	1	2	Максимальное время:	100	секунд	OK		
4	Кoeffициенты целевой функции	2	3	Предельное число итераций:	100		Отмена		
5	Ограничения	Кoeffициенты функций ограни		Относительная погрешность:	0,000001		Загрузить модель...		
6	Ограничение 1	1	1	Допустимое отклонение:	5	%	Сохранить модель...		
7	Ограничение 2	-6	7	Сходимость:	0,000		Справка		
8	Ограничение 3	1	1						
9									
10									
11									
12									
13									
14									
15									

Рис. 6.32. Диалоговое окно Параметры поиска решения

	A	B	C	D	E	F	G	H	
2		x1	x2	x3					
3	Переменные	0	0	2,5	Значение целевой функции				
4	Кoeffициенты целевой функции	2	3	5	12,5				
5	Ограничения	Кoeffициенты функций ограничений			Значения функций ограничений		Правая часть ограничений		
6	Ограничение 1	1	1	-1	-2,5	>=	-5	-2,5	
7	Ограничение 2	-6	7	-9	-22,5	<=	4	26,5	
8	Ограничение 3	1	1	4	10	=	10	5,732E-12	
9	Результаты поиска решения								
10	Решение найдено. Все ограничения и условия оптимальности выполнены.								
11									
12									
13	<input type="checkbox"/> Сохранить найденное решение								
14	<input type="checkbox"/> Восстановить исходные значения								
15									
16	<input type="button" value="OK"/> <input type="button" value="Отмена"/> <input type="button" value="Сохранить сценарий..."/> <input type="button" value="Справка"/>								
17									

Рис. 6.33. Успешное завершение решения задачи оптимизации

Диалоговое окно Результаты поиска решения сообщает о завершении поиска (см. рис. 6.33). То, что программа Поиск решения завершила работу, не означает, что она нашла оптимальное решение. Если оптимальное решение найдено, в диалоговом окне Результаты поиска решения должны присутствовать два ключевых предложения: Решение найдено и Все ограничения и условия оптимальности выполнены. Если хотя бы одного из этих предложений нет, программе не удалось оптимизировать модель. В таком случае следует сначала проверить правильность внесения данных в диалоговое окно Поиск решения, затем проверить табличную модель и наконец пересмотреть исходную формулировку задачи.

Если получено сообщение об успешном завершении поиска, можно или сохранить найденное решение, выбрав соответствующую опцию, или отбросить его, выбрав опцию Восстановить исходные значения. В результате ячейкам пере-

менных будут возвращены значения, которые в них находились до запуска программы Поиск решения. Существует возможность также получить три типа отчетов о решении. Каждый отчет выводится на новый лист рабочей книги.

6.3.2. Задачи, решаемые средством Поиск решения

Опишем четыре основных типа задач, которые можно решить с помощью средства "Поиск решения. Оптимизация при наличии ограничений — самый общий тип задачи оптимизации; другие типы задач появляются при ограничениях специального вида или их отсутствии. Эти задачи могут решаться и как задачи линейной оптимизации, и как задачи нелинейной оптимизации.

1. Поиск допустимого решения. Если не задавать целевую ячейку (в поле ввода Установить целевую ячейку в диалоговом окне Поиск решения), то средство Поиск решения остановит работу, найдя допустимое решение задачи, т.е. набор значений для изменяемых ячеек, которые удовлетворяют всем ограничениям. Если все функции ограничений линейные, то, установив флажок Линейная модель в диалоговом окне Параметры поиска решения, можно ускорить поиск допустимого решения.
2. Подбор параметров. Целевая ячейка не задается, указываются ограничения *только в виде равенств* или задается конкретное значение для целевой ячейки без определения каких-либо ограничений. В первом случае выполняется поиск тех значений изменяемых ячеек, которые удовлетворяют заданной системе ограничений, т.е., по сути, решается система уравнений, в которой неизвестными являются значения изменяемых ячеек. (Если некоторые ограничения заданы в виде неравенств, Поиск решения находит допустимое решение, определяемое заданной системой ограничений (см. задачу 1).) Во втором случае (когда задано конкретное значение целевой функции без указания ограничений) Поиск решения работает подобно средству Excel Подбор параметра, при этом используя другой алгоритм поиска. Кроме того, в отличие от средства Подбор параметра, Поиск решения может проводить подбор нескольких параметров, доставляющих заданное значение целевой функции.
3. Поиск безусловного оптимума — задача нахождения максимума или минимума целевой функции при отсутствии ограничений. Эта задача имеет смысл только в том случае, если целевая функция является нелинейной (по отношению к значениям изменяемых ячеек). В случае попытки поиска оптимума линейной целевой функции (без задания ограничений) будет выводиться сообщение о неограниченном решении. Если целевая функция имеет несколько максимумов или минимумов, то Поиск решения находит один из них (локальный оптимум), который может не совпадать с глобальным оптимумом. Какой конкретно будет найден локальный оптимум, зависит от начальных значений изменяемых ячеек.
4. Поиск оптимума при наличии ограничений. Наиболее общей задачей является *задача условной оптимизации*, когда заданы ограничения и адрес ячейки целевой функции, которую необходимо максимизировать или минимизировать. Если целевая функция и все ограничения линейны, то это задача линейной оптимизации. Решение этой задачи будет найдено быстрее, на-

дежнее и с более подробной дополнительной информацией, если в диалоговом окне Параметры поиска решения установлен флажок Линейная модель. В противном случае Поиск решения использует метод приведенного градиента. Если целевая функция имеет несколько оптимумов, которые удовлетворяют ограничениям, то Поиск решения найдет один из них (т.е. локальный оптимум), который может не быть глобальным. Какой конкретно будет найден локальный оптимум, зависит от начальных значений изменяемых ячеек.

6.3.3. Примеры применения средства Поиск решения

Рассмотрим два примера применения средства Поиск решения. Первый пример показывает решение системы линейных алгебраических уравнений на основе данных примера из раздела 6.1.5, в котором показан другой способ решения таких систем. Во втором примере показано, как на основании критерия % подобрать параметры вероятностного распределения.

Решение системы линейных алгебраических уравнений

Исходная табличная модель для этой задачи показана на рис. 6.34. В данной модели вычисления производятся только в столбце Е, где вычисляются значения левых частей уравнений (формулы показаны на рис. 6.34). Заполненное диалоговое окно Поиск решения для данной задачи представлено на рис. 6.35, а найденное решение — на рис. 6.36. Как видно на последнем рисунке, средством Поиск решения найдено точное решение системы.

E6		=СУММПРОИЗВ(\$A\$4:\$D\$4,A6:D6)				
	A	B	C	D	F	G
1	Решение системы линейных алгебраических уравнений					
2	Переменные					
3	x1	x2	x3	x4		
4	1	2	3	4		
5	Матрица коэффициентов системы				Значения левых частей уравнений	Вектор правых частей уравнений
6		1	1	4	22	15
7	-2	3	4	0	16	11
8	0	2	-5	1	-7	-5
9					21	16
10						
11	=СУММПРОИЗВ(\$A\$4:\$D\$4,A9:D9)					

Рис. 6.34. Исходная табличная модель для решения системы линейных алгебраических уравнений

Подбор параметров вероятностного распределения

На рис. 6.37 показаны исходные данные: выборка из генеральной совокупности, имеющей нормальное распределение с математическим ожиданием 1 и среднеквадратическим отклонением 2 (выборка создана с помощью средства Генерация случайных чисел из пакета анализа). Диапазон ячеек, содержащий выборочные значения, назван Выборка (это имя используется в формулах). Выборочные среднее и среднеквадратическое отклонения равны соответственно 0,91903 и 2,171256

(значения в ячейках H1 и H2). Границы интервалов карманов записаны в диапазоне C5:C14, а частоты (диапазон D5:D15) подсчитаны с помощью средства Гистограмма из пакета анализа. Ожидаемые частоты вычисляются по формулам, которые показаны на рис. 6.38. Формула вычисления значения критериальной статистики для критерия χ^2 (см. раздел 2,4.3) записана в ячейке G7. Отметим, что это формула массива, которая позволяет избежать промежуточных вычислений. Значение критериальной статистики для случая, когда в качестве математического ожидания и среднеквадратического отклонения распределения генеральной совокупности использованы соответствующие выборочные значения, показано на рис. 6.37.

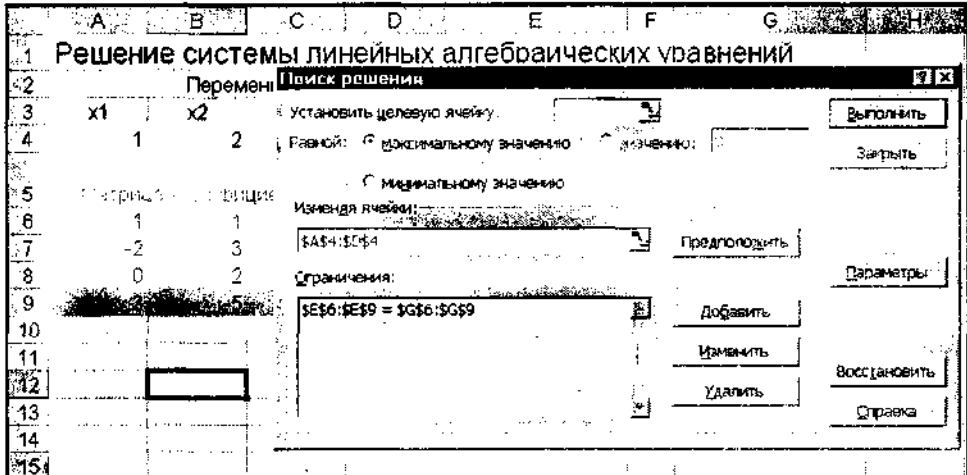


Рис. 6.35. Диалоговое окно Поиск решения для данной задачи

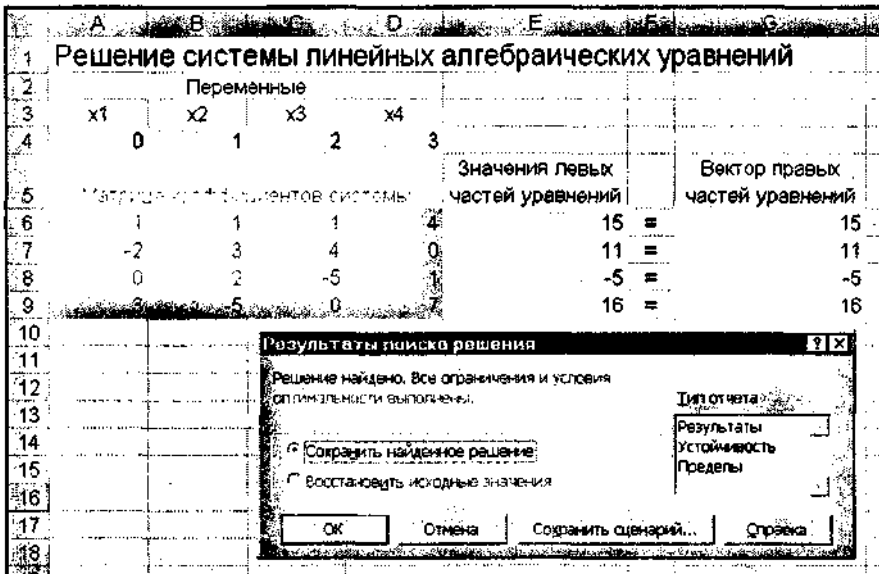


Рис. 6.36. Найденное решение системы линейных алгебраических уравнений

Изменяя значения математического ожидания и среднеквадратического отклонения, записанные в ячейках D1 и D2, с помощью средства Поиск решения попробуем уменьшить значение критериальной статистики, тем самым подобрав значения параметров распределения, которое лучше соответствует выборке. Диалоговое окно Поиск решения для этой задачи показано на рис. 6.39. В данном случае присутствует только одно ограничение — среднеквадратическое отклонение должно быть положительным. Но можно задать и другие ограничения, например, чтобы ожидаемая частота в крайних карманах имела значения не менее 1, как советуют некоторые статистические руководства.

Решение, полученное с помощью Поиск решения, показано на рис. 6.40. Полученные значения математического ожидания и среднеквадратического отклонения немного дальше от истинных, чем выборочные оценки, но значение критериальной статистики стало меньше, т.е. распределение с этими значениями математического ожидания и среднеквадратического отклонения лучше подходит для аппроксимации распределения выборки.

H7 {=СУММ(((D5:D15-E5:E15)^2)/E5:E15)}					
A	B	C	D	E	G
1	Выборка	Матем. ожидание	0,91903	Выборочное среднее	0,91903
2	0,399536	Станд. отклонение	2,171256	Выборочное станд. отклонение	2,171256
3	-1,55537				
4	1,486515	Карманы	Частота	Ожидаемая частота	
5	3,552947		-4	1	1,174027173
6	3,3967		-3	2	2,380015285
7	4,466266		-2	7	5,386966202
8	-3,36718		-1	9	9,898277311
9	0,531638		0	16	14,76558992
10	3,190045		1	18	17,88251211
11	-1,1734		2	14	17,58328951
12	-0,38041		3	16	14,03671278
13	-2,38086		4	9	9,097370348
14	-2,69382		5	5	4,786729563
15	-0,95526	Еще	3	3,008509799	
16	-0,54701				

Рис. 6.37. Исходные данные

E		G	H
1			
2			В = СРЗНАЧ(Выборка)
3			В = СТАНДОТКЛОН(Выборка)
4	Ожидаемая частота		
5	=НОРМРАСП(C5,\$D\$1,\$D\$2,1)*100		
6	=((НОРМРАСП(C6,\$D\$1,\$D\$2,1)-НОРМРАСП(C5,\$D\$1,\$D\$2,1))*100		
7	=((НОРМРАСП(C7,\$D\$1,\$D\$2,1)-НОРМРАСП(C6,\$D\$1,\$D\$2,1))*100		
8	=((НОРМРАСП(C8,\$D\$1,\$D\$2,1)-НОРМРАСП(C7,\$D\$1,\$D\$2,1))*100		
9	=((НОРМРАСП(C9,\$D\$1,\$D\$2,1)-НОРМРАСП(C8,\$D\$1,\$D\$2,1))*100		
10	=((НОРМРАСП(C10,\$D\$1,\$D\$2,1)-НОРМРАСП(C9,\$D\$1,\$D\$2,1))*100		
11	=((НОРМРАСП(C11,\$D\$1,\$D\$2,1)-НОРМРАСП(C10,\$D\$1,\$D\$2,1))*100		
12	=((НОРМРАСП(C12,\$D\$1,\$D\$2,1)-НОРМРАСП(C11,\$D\$1,\$D\$2,1))*100		
13	=((НОРМРАСП(C13,\$D\$1,\$D\$2,1)-НОРМРАСП(C12,\$D\$1,\$D\$2,1))*100		
14	=((НОРМРАСП(C14,\$D\$1,\$D\$2,1)-НОРМРАСП(C13,\$D\$1,\$D\$2,1))*100		
15	=((1-НОРМРАСП(C14,D1,D2,1))*100		
16			

Рис. 6.38. Формулы для вычислений

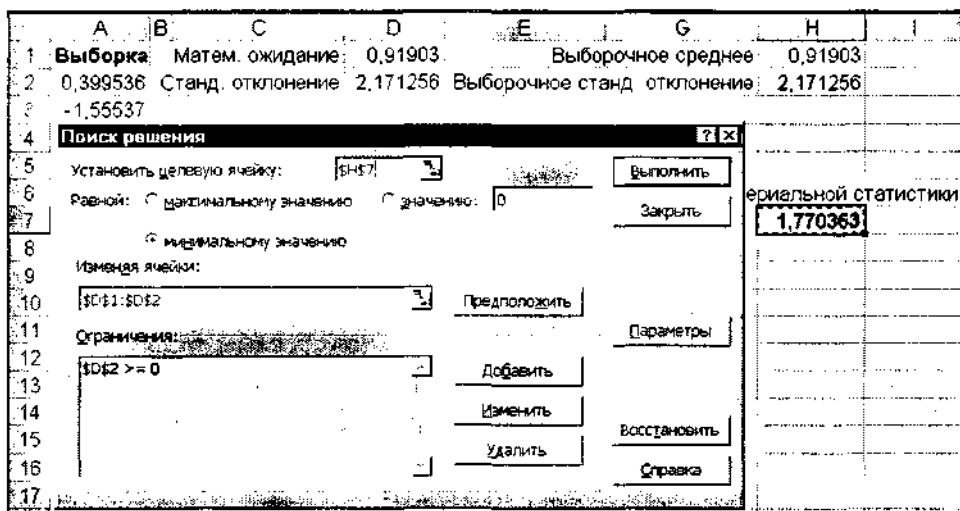


Рис. 6.39. Диалоговое окно Поиск решения

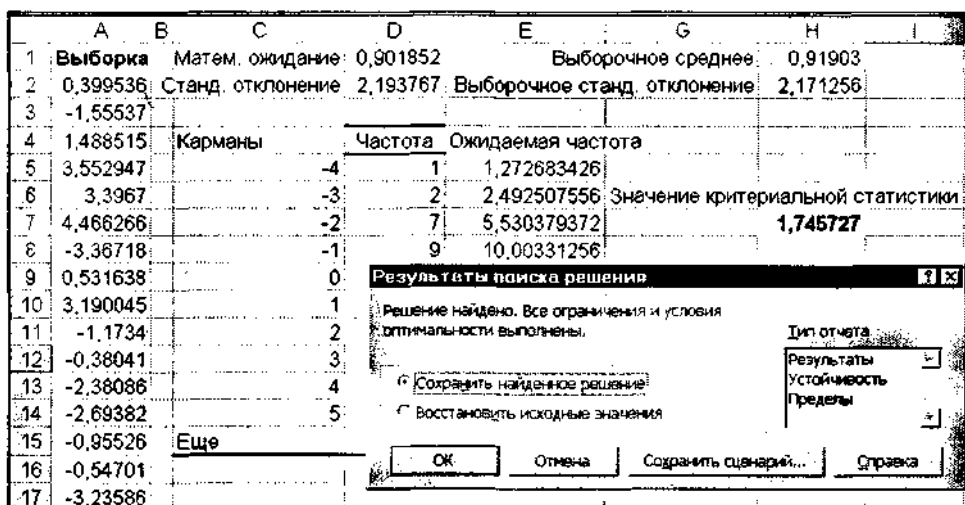


Рис. 6.40. Решение

Моделирование случайных величин

Моделирование случайных величин часто используется в статистическом анализе, хотя бы для построения тестовых выборок с заданными статистическими характеристиками, на основе которых можно проверить вычислительные алгоритмы методов математической статистики. В данной книге все примеры практической реализации описываемых методов иллюстрируются с использованием выборок, построенных в Excel. Но основное применение "искусственные" случайные величины находят в методах Монте-Карло и имитационном моделировании, где без таких случайных величин просто невозможно говорить об этих предметах.

Обычно рассмотрение темы моделирования случайных величин начинается с методов генерирования случайных чисел, имеющих равномерное распределение на интервале $[0, 1]$, так как эти величины являются основой для моделирования случайных величин, имеющих другие распределения. Мы не будем поднимать эту интересную тему, поскольку в Excel имеются готовые средства (функция СЛЧИС и средство Генерация случайных чисел) для создания последовательности равномерно распределенных случайных чисел. Рассмотрим способы моделирования произвольных случайных величин.

В Excel есть довольно много средств для генерирования значений случайных величин, имеющих различные распределения. Эти средства перечислены ниже. Но, конечно, встроенные средства Excel не обеспечивают моделирование вероятностных распределений "на все случаи жизни". Поэтому при необходимости генерирования случайных чисел, распределения которых нет в нижеприведенном списке, приходится вспоминать методы получения случайных значений, имеющиеся в арсенале теории вероятностей и математической статистики. Не вдаваясь в "глубокую" теорию, покажем применение *метода обратных функций*, *метода суперпозиций* и *метода отбора* для генерирования случайных чисел в Excel. В конце главы рассмотрим вопрос о моделировании зависимых случайных величин.

7.1. Средства Excel для генерирования случайных чисел

Перечислим имеющиеся в Excel средства для генерирования случайных чисел.

- Функция СЛЧИС, вычисляющая случайные числа, которые равномерно распределены на интервале $[0, 1]$ (см. раздел 4.13.1).

- Функция СЛУЧМЕЖДУ, генерирующая целочисленные значения, которые подчиняются дискретному равномерному распределению (см. раздел 4.13.2). (Функция доступна только тогда, когда подключена надстройка Пакет анализа.)
- Средство Генерация случайных чисел из надстройки Пакет анализа (см. раздел 5.3), предоставляющее возможность генерировать случайные числа, которые имеют следующие распределения.
 - Равномерное. Генерируется последовательность равномерно распределенных случайных чисел в заданном интервале.
 - Нормальное. Генерируется последовательность случайных чисел, подчиняющихся нормальному распределению. Задается математическое ожидание и среднеквадратическое отклонение.
 - Бернулли. Генерируется последовательность случайных чисел, принимающих только значение 0 или 1, в зависимости от заданной вероятности успеха (исхода "1"). (О распределении Бернулли речь идет в разделе 1.4.2.)
 - Биномиальное. Генерируется последовательность случайных чисел, равное количеству исходов "1" в n независимых испытаниях. В результате каждого из них с вероятностью p может произойти исход "1" и с вероятностью $(1 - p)$ — исход "0" (см. раздел 1.4.3). Здесь необходимо задать число испытаний n и вероятность p .
 - Пуассона. Генерируется последовательность случайных чисел, подчиняющихся распределению Пуассона с заданным параметром X . (О распределении Пуассона речь идет в разделе 1.4.4.)
 - Дискретное. Генерируется последовательность случайных чисел, подчиняющихся заданному дискретному распределению. Для задания этого распределения необходимо указать диапазон ячеек, состоящий из двух столбцов: в первом столбце содержатся значения, а во втором — вероятности каждого значения.

Между способами вычисления случайных чисел, полученных с помощью функции СЛЧИС (соответствующие формулы приведены в следующих разделах) и с помощью средства Генерация случайных чисел, в частности равномерно распределенных на интервале $[0, 1]$, имеются существенные различия. Первое различие заключается в том, что функцию СЛЧИС можно непосредственно использовать в формулах (в том числе в формулах массивов) как аргумент формулы или другой функции, тогда как для того, чтобы использовать в формулах случайные числа, полученные с помощью средства Генерация случайных чисел, сначала необходимо их получить, т.е. записать в отдельном диапазоне ячеек, и только затем использовать в формулах.

Второе отличие состоит в том, что формулы, содержащие функцию СЛЧИС, пересчитываются при каждом пересчете рабочего листа (например, при любом вводе значения в ячейку или при удалении чего-либо, или при нажатии клавиши <F9>), а значения, полученные с помощью средства Генерация случайных чисел, фиксированы — при необходимости получения новой выборки на месте старой, следует еще раз вызвать и применить это средство. Свойство

"изменчивости" функции СЛЧИС полезно, например, в имитационном моделировании. Однако в других случаях оно может сильно замедлять работу в Excel или быть просто излишним. Чтобы зафиксировать значения, вычисляемые с помощью функции СЛЧИС, надо выделить диапазон ячеек, содержащий эти значения, и скопировать его (команда Правка^Копировать). Затем, не отменяя выделения диапазона, следует выполнить команду Правка^Специальная вставка, в открывшемся диалоговом окне Специальная вставка установить переключатель Значения и щелкнуть на кнопке ОК. В ячейки выделенного диапазона вместо формул будут записаны числовые значения.

Покажем, как можно использовать пересчет функции СЛЧИС для получения на основе одной выборки результатов нескольких экспериментов, когда "выходом" эксперимента является сама выборка. На рис. 7.1 в столбце А показана выборка объемом 15 значений, полученная по формуле массива {=СЛЧИС()}. Пусть по выборочным значениям в столбце В вычисляются среднее, выборочная дисперсия, минимальное и максимальное значения по стандартным формулам. (Эти статистические характеристики в данном случае выбраны произвольно; в зависимости от конкретных целей могут вычисляться другие величины, например критерийные статистики или интервальные оценки каких-либо параметров.)

	A2				
		B	C	D	E
1	Выборка	Среднее	Дисперсия	Минимум	Максимум
2	0,684621	0,564033	0,09247015	0,098473	0,9405703
3	0,876142				
4	0,801418				
5	0,94057				
6	0,552437				
7	0,695795				
8	0,612567				
9	0,234976				
10	0,911106				
11	0,180257				
12	0,512054				
13	0,098473				
14	0,318471				
15	0,911406				
16	0,130205				
17					

Рис. 7.1. Выборка и ее характеристики

Если в данной ситуации нажать клавишу <F9>, выборочные значения изменятся и соответственно изменятся значения в столбце В. Таким образом, получается новая выборка с тем же распределением. Осталось зафиксировать новые значения характеристик. Для этого их можно скопировать (как значения, а не формулы!) в отдельный диапазон ячеек. Но это неудобный прием, если надо провести несколько экспериментов с одной и той же выборкой.

Покажем, как можно выполнить сразу столько экспериментов, сколько необходимо, и заодно сразу получить все значения характеристик для каждого эксперимента. Для этого можно использовать таблицы подстановки Excel. В данном случае целесообразно использовать таблицу подстановки с одним входом. Для

создания такой таблицы введем сначала последовательность целых чисел от 1 до числа, задающего количество экспериментов. В нашем примере введем числа от 1 до 12 в столбце D, как показано на рис. 7.2¹. В ячейку E3 введем формулу =B2, в ячейку F3 — формулу =B4, в ячейку G3 — =B6 и в ячейку H3 — =B8. Эти формулы указывают, какие характеристики будут вычисляться в таблице подстановки. Для пояснения можно добавить заголовки столбцов Среднее, Дисперсия и т.д., как показано на рис. 7.2.

E3				=B2				
	A	B	C	D	E	F	G	H
1	Выборка	Среднее			Таблица подстановки			
2	0,436749	0,604042			Среднее	Дисперсия	Минимум	Максимум
3	0,524251	Дисперсия			0,604042	0,06786598	0,069171	0,9455233
4	0,882643	0,067866			1			
5	0,649284	Минимум			2			
6	0,834089	0,069171			3			
7	0,795292	Максимум			4			
8	0,770848	0,945523			5			
9	0,711879				6			
10	0,573031				7			
11	0,945523				8			
12	0,196071				9			
13	0,820599				10			
14	0,310545				11			
15	0,069171				12			
16	0,540658							
17								

Рис. 7.2. Подготовка таблицы подстановки

Далее следует выделить диапазон ячеек D3:H15 и выполнить команду Данные^{1^} Таблица подстановки, в результате чего откроется диалоговое окно Таблица подстановки (рис 7.3). В этом диалоговом окне поле Подставлять значения по столбцам в оставим пустым (оно заполняется, если для таблицы подстановки числовые значения записаны в строку, а формулы — в столбец). В поле Подставлять значения по строкам в введем адрес любой пустой ячейки (на рис. 7.3 показан адрес ячейки И). В данном случае, в отличие от "настоящих" таблиц постановки, числа в столбце D не участвуют в вычислениях; здесь акт их подстановки в указанную ячейку является "спусковым механизмом" для нового пересчета формул, содержащих функцию СЛЧИС. После щелчка на кнопке ОК в окне Таблица подстановки выделенная область будет заполнена результатами расчетов, как показано на рис. 7.4.

Результаты расчетов в таблице подстановки подвержены изменению (т.е. автоматически пересчитываются), например, при нажатии клавиши <F9>. Чтобы зафиксировать эти значения, следует или преобразовать их в значения с помощью диалогового окна Специальная вставка, как рассказано выше, либо скопировать и вставить их как значения в новый диапазон ячеек (а в таблице подстановки можно продолжать эксперименты).

¹ В принципе, эти числа могут быть любыми, в том числе равными, дробными или отрицательными, что не влияет на дальнейшие вычисления. Но в виде натуральных чисел они могут нести смысловую нагрузку как порядковые номера экспериментов.

	A	B	C	D	E	F	G	H	I	J
1	Выборка	Среднее								
2	0,436749	0,604042								
3	0,524251	Дисперсия								
4	0,882643	0,067866								
5	0,649284	Минимум								
6	0,834089	0,069171								
7	0,795292	Максимум								
8	0,770848	0,945523								
9	0,711879									
10	0,573031									
11	0,945523									
12	0,196071									
13	0,820599									
14	0,310545									
15	0,069171									
16	0,540658									
17										

	Среднее	Дисперсия	Минимум	Максимум
1	0,604042	0,06786598	0,069171	0,9455233
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				

Подставлять значения по столбцам в:	
Подставлять значения по строкам в:	\$F\$1
<input type="button" value="ОК"/> <input type="button" value="Отмена"/>	

Рис. 7.3. Диалоговое окно Таблица подстановки

	A	B	C	D	E	F	G	H
1	Выборка	Среднее						
2	0,722519	0,614922						
3	0,075301	Дисперсия						
4	0,742908	0,062741						
5	0,309161	Минимум						
6	0,480562	0,075301						
7	0,813407	Максимум						
8	0,677483	0,957561						
9	0,776775							
10	0,915317							
11	0,536159							
12	0,828199							
13	0,957561							
14	0,419216							
15	0,331985							
16	0,63728							
17								

	Среднее	Дисперсия	Минимум	Максимум
1	0,614922	0,06274146	0,075301	0,9575611
2				
3				
4				
5				
6				
7				
8				
9				
10				
11				
12				

Рис. 7.4. Результаты расчетов

Функцию СЛЧИС легко применить для моделирования случайных величин, равномерно распределенных на произвольном интервале $[a, b]$. Для этого необходимо использовать формулу

$$=(b-a)*\text{СЛЧИС}()+a,$$

где вместо a и b подставляются конкретные числа или ссылки на ячейки, содержащие эти числа.

Также отметим, что формула $=\text{ЦЕЛОЕ}((b-a)*\text{СЛЧИС}()+a)$ генерирует целочисленные значения, равномерно распределенные на интервале $[a, b - 1]$, т.е. является аналогом функции СЛУЧМЕЖДУ.

7.2. Метод обратных функций моделирования случайных величин

В основе этого метода лежит известный математический факт, что если $G(x)$ — функция, обратная к функции распределения $F(x)$, то случайная величина $Y = G(X)$, где случайная величина X распределена равномерно на интервале $[0, 1]$, имеет функцию распределения $F(x)$ (см. раздел 1.3). В Excel есть несколько функций, возвращающих значения обратных функций для различных распределений. Это следующие функции (см. раздел 4.7).

- **FPACнОВР**. Возвращает обратное значение для **F**-распределения.
- **БЕТАОБР**. Вычисляет значение функции, обратной к функции бета-распределения.
- **ГАММАОБР**. Вычисляет значение функции, обратной к функции гамма-распределения.
- **ЛОГНОРМОБР**. Вычисляет значение функции, обратной к функции логарифмически нормального распределения.
- **НОРМОБР**. Вычисляет значение функции, обратной к функции нормального распределения.
- **НОРМСТОБР**. Вычисляет значение функции, обратной к функции стандартного нормального распределения.
- **СТЮДРАСПОБР**. Вычисляет значение функции, обратной к функции распределения Стьюдента.
- **ХИ2ОБР**. Вычисляет значение функции, обратной к функции распределения χ^2 .

Таким образом, формула **=ФУНКЦИЯ(СЛЧИС();...)**, где **ФУНКЦИЯ** обозначает одну из вышеперечисленных функций с соответствующими аргументами, будет генерировать последовательность случайных чисел, которые имеют распределение, определяемое данной функцией². Этим способом можно генерировать случайные величины, имеющие

- \wedge -распределение (распределение Снедекора);
- бета-распределение;
- гамма-распределение;
- логарифмически нормальное распределение;
- нормальное распределение;
- распределение Стьюдента;
- распределение χ^2 .

¹ Некоторые из перечисленных функций являются обратными не к функции распределения $F(x)$, а к функции $1 - F(x)$, поэтому, строго говоря, в таких функциях вместо аргумента **СЛЧИС()** должен стоять аргумент $1 - \text{СЛЧИС}()$. Но поскольку случайные величины X и $1 - X$ имеют одинаковые распределения, если X равномерно распределена на интервале $[0, 1]$, то приведенная формула справедлива для любых обратных функций.

Моделирование случайных величин, имеющих распределение Стьюдента, требует пояснения, поскольку функция Excel СТЬЮДРАСПОБР не возвращает отрицательных значений, — она предназначена для использования в статистических критериях для вычисления критических значений, но не для генерирования случайных чисел. Однако, поскольку это распределение симметрично относительно нуля, случайная величина X , имеющая распределение Стьюдента, с вероятностью 0,5 может принимать отрицательные значения и с такой же вероятностью — положительные. Исходя из этого замечания для генерирования случайных чисел с данным распределением можно применить формулу $\text{=ЕСЛИ(СЛЧИС()<0,5;-СТЬЮДРАСПОБР(СЛЧИС());СТЬЮДРАСПОБР(СЛЧИС());K)}$. Здесь аргумент K задает число степеней свободы распределения Стьюдента. Формулу можно применять как формулу массива для генерирования выборки нужного размера. Эта формула использована в примере из раздела 9.2.1.

Из вышеприведенного списка только нормально распределенные случайные числа можно получить с помощью средства Генерация случайных чисел. Таким образом, метод обратных функций с использованием встроенных функций Excel позволяет моделировать широкий спектр вероятностных распределений, особенно с учетом того, что многие другие распределения, не вошедшие в вышеприведенный список, являются частными случаями либо бета-распределения (например, распределение арксинуса, треугольное и даже равномерное), либо гамма-распределения (например, распределение Эрланга и показательное распределение).

Если необходимо моделировать случайную величину, распределения которой нет в приведенном выше списке, но известна функция, обратная к ее функции распределения, то используют формулу, вычисляющую эту обратную функцию с аргументом СЛЧИС. Например, известно, что функция показательного распределения имеет вид $F(x) = 1 - e^{-\lambda x}$ ($x \geq 0$), где X — параметр распределения, $X > 0$. Обратная функция,

как нетрудно показать, определяется формулой $G(x) = -\frac{1}{\lambda} \ln(1-x)$. Поэтому для генерирования случайных чисел, имеющих показательное распределение, можно использовать формулу =-1*M(СЛЧИС())/A1 , если значение λ записано в ячейке A1.

Таким образом, если известна функция, обратная к функции распределения, то моделирование *непрерывных* случайных величин не вызывает особых затруднений. Моделирование *дискретных* случайных величин методом обратных функций вызывает определенные сложности, связанные с тем, что для дискретных случайных величин функция распределения имеет ступенчатый вид и поэтому обратная функция определяется неоднозначно. Существует несколько подходов к построению обратных *функций* дискретных распределений, и хотя по сути они достаточно просты, на практике использовать их неудобно. Excel позволяет моделировать дискретные случайные величины без непосредственного построения обратной функции. (Не забывая, что средство Генерация случайных чисел также позволяет моделировать любые дискретные величины, но здесь мы обойдемся без этого средства.)

На рис. 7.5 показана таблица, содержащая значения, вероятности этих значений и значения функции распределения. Чтобы моделировать случайную величину с таким распределением, надо выполнить такие действия. Перед столбцом, содержащим значения случайной величины, вставить еще один столбец, содержащий значения функции распределения, как показано на рис. 7.6. Обращаем внимание, что первое значение в этом столбце равно 0. Данные в столбцах C и D в дальнейшем не используются.

	A	B	C	D	E
1		Значения	Вероятность	Функция распределения	
2		1,5	0,24	0,24	
3		2	0,09	0,33	
4		2,5	0,2	0,53	
5		4	0,08	0,61	
6		4,5	0,06	0,67	
7		5	0,09	0,76	
8		6	0,24	1	
9					
10					

Рис. 7.5. Распределение дискретной случайной величины

	A	B	C	D	E
1	Функция распределения	Значения	Вероятность	Функция распределения	
2		0	1,5	0,24	0,24
3	0,24		2	0,09	0,33
4	0,33		2,5	0,2	0,53
5	0,53		4	0,08	0,61
6	0,61		4,5	0,06	0,67
7	0,67		5	0,09	0,76
8	0,76		6	0,24	1
9	1				
10					

Рис. 7.6. Подготовка к моделированию

Для генерирования случайных чисел в данном случае используется формула

$$=ВПР(СЛЧИС();A2:B8;2).$$

Функция ВПР в первом столбце таблицы, задаваемой вторым аргументом A2:B8, ищет совпадения со значением первого аргумента СЛЧИС(). При наличии такого совпадения функция возвращает значение из второго столбца (номер столбца задает третий аргумент 2) и строки, в которой было обнаружено совпадение значений. Если точного совпадения нет, то в качестве искомого берется ближайшее значение, не превосходящее значение первого аргумента. Так работает эта функция, если не задан ее четвертый (необязательный) аргумент. Приведенную формулу можно применять как формулу массива для генерирования не одного, а нескольких случайных чисел. На рис. 7.7 в столбце F показан массив сгенерированных значений. Эти значения пересчитываются при нажатии клавиши <F9>, поэтому их можно использовать для получения нескольких выборок, имеющих одинаковые распределения.

На практике метод обратных функций используется в основном тогда, когда известно аналитическое выражение функции, обратной к функции распределения. Но в Excel есть средства, которые позволяют генерировать случайные числа без использования явного вида обратной функции, что показывает приведенный пример моделирования дискретной случайной величины. Для моделирования непрерывных случайных величин можно использовать средства Подбор параметра и Поиск решения для получения значений обратной функции путем решения уравнения $F(x) = 4>$ где ξ , — заданное значение случайной величины, имеющей равномерное распределение на интервале $[0, 1]$. Покажем, как это делается, на примере

генерирования нормально распределенных случайных чисел, поскольку считается, что нормально распределенные случайные числа методом обратных функций генерировать весьма сложно и такие числа обычно генерируют с помощью других методов.

F2		{=ВПР(СЛЧИС();A2:B8;2)}					
	A	B	C	D	E	F	G
1	Функция	Значения	Вероятность	Функция	Случайные		
2	распределения			распределения	числа		
2		0	1,5	0,24		6	
3		0,24	2	0,09		5	
4		0,33	2,5	0,2		5	
5		0,53	4	0,08		6	
6		0,61	4,5	0,06		4,5	
7		0,67	5	0,09		4,5	
8		0,76	6	0,24		1,5	
9		1		1		2	
10						2,5	
11						5	
12						2,5	
13						4	
14						2	
15						6	
16							

Рис. 7.7. Генерирование случайных чисел

Создадим таблицу, показанную на рис. 7.8. В столбце А введем столько чисел, сколько их должно быть в будущей выборке. Эти числа могут быть произвольными. Единственное ограничение, которое на них накладывается, заключается в том, чтобы функция распределения, вычисляемая на их основе, не принимала крайних значений 0 и 1, поскольку это затруднит работу средства Поиск решения. В столбце В вычисляются значения функции распределения, в данном случае по формуле $=\text{НОРМСТРАСП}(A2)$, которая записана в ячейке В2 и затем скопирована вниз до конца интервала. В столбце С с помощью формулы массива $\{\text{СЛЧИС}()\}$ сгенерированы случайные числа, имеющие равномерное распределение на интервале $[0, 1]$. Затем формула преобразуется в значения с помощью диалогового окна Специальная вставка.

Далее применяется средство Поиск решения, диалоговое окно которого показано на рис. 7.9. В данном случае с помощью этого средства вычисляются корни уравнений $F(x) = 4$, значения b , которых записаны в столбце С, а значения корней x будут записаны в столбце А (об использовании средства Поиск решения для решения уравнений речь идет в разделе 6.3). Одновременно решается столько уравнений, сколько необходимо сгенерировать выборочных значений. В диалоговом окне Поиск решения целевая ячейка не задается, в качестве изменяемых ячеек указываются все ячейки столбца А, в которые введены числа. Ограничения в данном случае задаются в виде одного равенства $B2:B16 = C2:C16$. После щелчка на кнопке Выполнить Excel после некоторого времени "раздумий", длительность которого зависит от количества решаемых уравнений, найдет корни всех уравнений и тем самым сгенерирует случайные числа. Результат вычислений показан на рис. 7.10.

Подобным образом можно сгенерировать значения любой случайной величины с известной функцией распределения, если сама функция распределения достаточно гладкая, поскольку Поиск решения для решения уравнений использует градиентный метод.

B2		=НОРМСТРАСП(A2)	
A		B	C
Случайные		Функция	Равномерно
числа		распределения	распределенные
1			числа
2	1	0,84134474	0,772499434
3	1	0,84134474	0,927763076
4	1	0,84134474	0,164312844
5	1	0,84134474	0,147739842
6	1	0,84134474	0,668357236
7	1	0,84134474	0,116124329
8	1	0,84134474	0,605122167
9	1	0,84134474	0,619325302
10	1	0,84134474	0,879585378
11	1	0,84134474	0,474866944
12	1	0,84134474	0,014485647
13	1	0,84134474	0,891356599
14	1	0,84134474	0,483318402
15	1	0,84134474	0,338995213
16	1	0,84134474	0,596897227
17			

Рис. 7.8. Подготовка к моделированию

A	B	C	D	E	F	G
	Случайные	Функция	Равномерно			
	числа	распределения	распределенные			
1			числа			
2	1	0,84134474	0,772499434			
3	1	0,84134474	0,927763076			
4	1					
5	1					
6	1					
7	1					
8	1					
9	1					
10	1					
11	1					
12	1					
13	1					
14	1					
15	1					
16	1					
17						
18						

Поиск решения

Установить целевую ячейку:

к: ☐ максимальному значению ☐ значению: 0 ☐ минимальному значению

Изменяя ячейки:

Ограничения:

Предположить

Добавить

Изменить

Удалить

Выполнить

Закрыть

Параметры

Восстановить

Справка

Рис. 7.9. Диалоговое окно Поиск решения

7.3. Метод суперпозиций

Данный метод генерирования значений случайной величины X применяется тогда, когда ее функцию распределения $F(x)$ можно представить в виде суммы

$$F(x) = \sum_{i=1}^n c_i F_i(x), \text{ где все } F_i(x) \text{ — та же функция распределения, а все коэффициенты } c_i \text{ — различные значения.}$$

енты $c_k > 0$, при этом $c_1 + c_2 + \dots + c_m = 1$. (Такая случайная величина X называется *смесью случайных величин*.) Коэффициенты c_k можно рассматривать как вероятности, задающие распределение дискретной случайной величины Y , которая принимает целочисленные значения k с вероятностью c_k . Доказано [16, с. 64], что если в соответствии с распределением величины Y выбирать номер k , а затем из уравнения $F_k(X) = \xi$, где ξ — значение случайной величины, имеющей равномерное распределение на интервале $[0, 1]$, определить X , то случайная величина X

подчиняется вероятностному закону с функцией распределения $F(x) = \sum_{k=1}^m c_k F_k(x)$.

	А	В	С
1	Случайные числа	Функция распределения	Равномерно распределенные числа
2	0,74710319	0,772499434	0,772499434
3	1,45933193	0,927763076	0,927763076
4	-0,9768858	0,164312844	0,164312844
5	-1,0461763	0,147739842	0,147739842
6	0,43538159	0,668357236	0,668357236
7	-1,1945867	0,116124329	0,116124329
8	0,26662806	0,605122167	0,605122167
9	0,30370944	0,619325302	0,619325302
10	1,17291691	0,879585378	0,879585378
11	-0,0630408	0,474866944	0,474866944
12	-2,1838899	0,014485118	0,014485647
13	1,18179593	0,881356599	0,881356599
14	-0,0418266	0,483318402	0,483318402
15	-0,415207	0,338995213	0,338995213
16	0,24532416	0,596897227	0,596897227
17			

Рис. 7.10. Сгенерированные случайные числа

На основе этого утверждения можно построить следующую схему вычисления значений случайной величины X . Пусть имеются два сгенерированных независимых случайных числа ξ_1 и ξ_2 , равномерно распределенных на интервале $[0, 1]$. Значение x случайной величины X вычисляется по формуле

$$x = \begin{cases} G_1(\xi_2), & \text{если } \xi_1 < c_1, \\ G_2(\xi_2), & \text{если } c_1 < \xi_1 < c_1 + c_2, \\ \vdots \\ G_k(\xi_2), & \text{если } \sum_{i=1}^{k-1} c_i < \xi_1 < \sum_{i=1}^k c_i, \\ \vdots \\ G_m(\xi_2), & \text{если } \sum_{i=1}^{m-1} c_i < \xi_1 < 1. \end{cases}$$

где G_k — функции, обратные к функциям распределения F_k (т.е. здесь используется метод обратных функций).

Покажем, как эту формулу можно реализовать в Excel. Сначала рассмотрим простой случай, когда $m = 2$. Пусть $F(x) = \frac{2}{5}F_1(x) + \frac{3}{5}F_2(x)$, $F_1(x)$ — функция распределения показательного закона с параметром $\lambda = 2$ (для этого распределения обратная функция имеет вид $G(x) = -\frac{1}{\lambda} \ln(1-x)$), $F_2(x)$ — функция распределения нормального закона с параметрами $m = 1$ и $\sigma = 2$ (здесь для вычисления обратной функции будем использовать функцию Excel НОРМОБР). Для генерирования значений случайной величины, имеющей функцию распределения $F(x) = \frac{2}{5}F_1(x) + \frac{3}{5}F_2(x)$, надо применить формулу

$$=ЕСЛИ(СЛЧИС()<2/5; -LN(СЛЧИС())/2; НОРМОБР(СЛЧИС();1;2)).$$

Ее можно использовать как формулу массива, сгенерировав при этом столько значений, сколько необходимо. Также отметим, что эти значения будут пересчитываться при нажатии клавиши <F9>. Таким образом каждый раз можно получать новую выборку.

Если $m \geq 3$, простой формулы для генерирования случайных чисел не существует³. Для случайного выбора обратной функции G_i , которая даст очередное значение случайной величины в соответствии с приведенной выше формулой, в Excel можно использовать функцию ВПР так, как при моделировании дискретных случайных величин (см. предыдущий раздел). Пусть, например,

$$F(x) = 0,1F_1(x) + 0,3F_2(x) + 0,2F_3(x) + 0,1F_4(x) + 0,3F_5(x),$$

где F_1 — функция бета-распределения с параметрами 1 и 2, F_2 — функция гамма-распределения с параметрами 2 и 3, F_3 — функция логарифмически нормального распределения с параметрами 1 и 1, F_4 — функция стандартного нормального распределения и F_5 — функция распределения Стьюдента с 10-ю степенями свободы.

На рис. 7.11 показана подготовительная таблица, содержащая коэффициенты, их частичные суммы, а также формулы с использованием обратных функций для вычисления случайных чисел. В ячейку D2 введем формулу =ВПР(СЛЧИС();\$A\$8:\$B\$12;2) и скопируем ее вниз столько раз, сколько необходимо. (Вводить эту формулу как формулу массива в данном случае нецелесообразно по причинам, которые будут понятны позже.) В результате получим случайные числа, показанные на рис. 7.12 в столбце Случайные числа.

Нетрудно заметить одинаковые значения среди сгенерированных чисел. Это следствие того, что случайные числа берутся из таблицы Обратные функции, которая, хотя и пересчитывается при каждом вычислении, имеет конечный набор значений. Чтобы обойти это препятствие и получить полноценную выборку, некоторые действия необходимо выполнить вручную. После каждого пересчета таблиц Обратные функции и Случайные числа, выполняемого с помощью клавиши <F9>.

Нельзя использовать вложенные функции ЕСЛИ для проверки значения, возвращаемого функцией СЛЧИС, поскольку для проверки более двух условий необходимо несколько "экземпляров" этого значения. Но если для получения этих "экземпляров" опять использовать функцию СЛЧИС, она даст другое значение. Если в формуле несколько раз встречается функция СЛЧИС, то она каждый раз генерирует новое значение. Например, формула =СЛЧИС()-СЛЧИС() никогда не возвратит нулевого значения.

	A	B	C
1	Номер функции	Коэффициенты	
2	1	0,1	
3	2	0,3	
4	3	0,2	
5	4	0,1	
6	5	0,3	
7	Суммы коэффициентов	Обратные функции	
8	0	0,659283638	=БЕТАОБР(СЛЧИС();1;2)
9	0,1	0,604795787	=ГАММАОБР(СЛЧИС();2;3)
10	0,4	2,952380181	=ЛОГНОРМОБР(СЛЧИС();1;1)
11	0,6	-1,119604477	=НОРМСТОБР(СЛЧИС())
12	0,7	1,840357982	=СТЮДРАСПОБР(СЛЧИС();10)
13			

Рис. 7.11. Подготовительная таблица

D2	=ВПР(СЛЧИС();\$A\$8:\$B\$12;1)			
	A	B	C	D
1	Номер функции	Коэффициенты	Случайные числа	
2	1	0,1	13,82733899	
3	2	0,3	2,758636443	
4	3	0,2	0,256623745	
5	4	0,1	0,256623745	
6	5	0,3	2,298438631	
7	Суммы коэффициентов	Обратные функции	2,291142896	
8	0	0,256623745	13,82733899	
9	0,1	13,82733899	0,256623745	
10	0,4	2,758636443	2,758636443	
11	0,6	2,291142896	2,291142896	
12	0,7	2,298438631	0,256623745	
13			13,82733899	
14			13,82733899	
15			2,758636443	
16			2,298438631	
17			13,82733899	
18				

Рис. 7.12. Сгенерированные случайные числа

необходимо поочередно зафиксировать числа (преобразовать формулу в значение) в столбце Случайные числа. (Из-за этого действия формулу в столбце Случайные числа не следует вводить как формулу массива, так как формула массива не позволяет манипулировать отдельными значениями.) Чтобы упростить это действие, можно написать простой макрос, который будет выполнять следующие действия. Пусть выделена отдельная ячейка. Сначала выполняется копирование содержимого ячейки любым способом (с помощью шелчка на кнопке Копировать стандартной панели инструментов или команды Правка^ Копировать). Затем по команде Правка^ОСпециальная вставка открывается одноименное диалоговое окно, в котором устанавливается переключатель Значения. Щелчок на кнопке ОК этого окна — последнее действие, которое надо записать в макрос. Перед началом записи макроса (команда Сервис=>Макрос^Начать запись) в окне Запись

макроса рекомендуем задать комбинацию клавиш, с помощью которой будет выполняться макрос. После щелчка на кнопке ОК в окне Запись макроса открывается панель Остановить запись, где перед записью действий *обязательно* надо щелкнуть на кнопке Относительные ссылки. Если использовать описанный макрос, то поочередное фиксирование значений выборки средних размеров (порядка 100 чисел) займет всего пару минут или чуть больше (в зависимости от скорости нажатия клавиш). На рис. 7.13 показана окончательная выборка.

D2 98873876857492			
	A	B	C
1	Номер функции	Коэффициенты	Случайные числа
2	1	0,1	8,988738769
3	2	0,3	0,124770164
4	3	0,2	0,053227041
5	4	0,1	0,545422836
6	5	0,3	4,817505503
7	Суммы коэффициентов	Обратные функции	0,00330374
8	0	0,105375767	5,233500815
9	0,1	0,989093234	2,111419235
10	0,4	2,885922244	-1,338435583
11	0,6	0,95239985	0,848475565
12	0,7	1,220395234	1,539060577
13			13,64303898
14			4,125713531
15			0,047722324
16			5,609772415
17			0,290007165
18			

Рис. 7.13. Окончательно сформированная выборка

7.4. Метод отбора

Этот метод применяется для моделирования непрерывных случайных величин, которые имеют сложное распределение на *конечном интервале* и для которых известно аналитическое выражение плотности вероятностей $p(x)$. Метод используется в основном тогда, когда другие методы моделирования неприемлемы. Рассмотрим простейший вариант метода отбора, который также называют *методом Неймана* (по имени его разработчика)⁴.

На рис. 7.14 показан график плотности вероятности $y = p(x)$ некоторой случайной величины X , распределенной на интервале $[a, b]$, и график функции $y = c$, которая мажорирует плотность $p(x)$. Пусть ξ и Γ — случайные величины, равномерно распределенные на интервалах $[a, b]$ и $[0, c]$ соответственно. Доказано [16, с. 76], что случайная величина X , определяемая условием $X = \xi$, если $\Gamma < p(\xi)$, имеет распределение с плотностью вероятностей $p(x)$. Другими словами, если двумерная случайная величина (ξ, Γ) , равномерно распределенная в прямоугольнике $a < x < b$, $0 < y < c$, попадает в область, лежащую ниже графика $y = p(x)$, то принимается, что $X = \xi$ (рис. 7.14). Величина c обычно

⁴ Иногда все методы отбора называют методами Неймана.

берется равной максимуму функции $p(x)$, но, если максимум неизвестен или сложно найти его точное значение, величина c берется заведомо большей, чем максимальное значение $p(x)$.

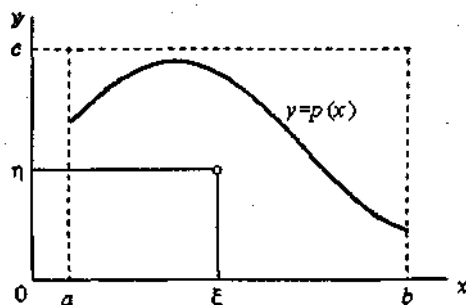


Рис. 7.14. Поясняющий рисунок к описанию метода отбора

На этом основании построен метод отбора: генерируются два независимых случайных числа ξ_1 и ξ_2 , равномерно распределенных на интервале $[0, 1]$, и вычисляются значения $x = a + (b - a)\xi_1$ и $y = c\xi_2$. Если выполняется неравенство $y < p(x)$, то за значение случайной величины X принимается значение x . В противном случае пара чисел ξ_1 и ξ_2 отбрасывается и генерируется новая, для которой выполняется аналогичная проверка.

Покажем реализацию этого алгоритма на примере моделирования случайной величины X , имеющей распределение на интервале $[-1, 1]$ с плотностью вероятности $p(x) = \frac{2}{\pi} \sqrt{1-x^2}$. Здесь по сгенерированным случайным числам ξ_1 и ξ_2 , рав-

номерно распределенным на интервале $[0, 1]$, проверяется неравенство $2\xi_2/\pi < p(2\xi_1 - 1)$, которое можно преобразовать в эквивалентное неравенство $\xi_2^2 < 1 - (2\xi_1 - 1)^2$. Если это неравенство выполняется, то за значение случайной величины X принимается число $2\xi_1 - 1$. На рис. 7.15 показан рабочий лист, в столбцах А и В которого с помощью формул =СЛЧИС() сгенерированы равномерно распределенные числа ξ_1 и ξ_2 . В столбце С по формуле

$$=ЕСЛИ(В2^2 < 1 - (2*A2-1)^2; 2*A2-1; "М"),$$

записанной в ячейке С2 и скопированной вниз, вычисляются значения случайной величины X . Если определяющее неравенство не выполняется, в ячейку столбца С записывается буква М (можно записать любое значение, показывающее, что в данной ячейке нет выборочного значения). Далее остается удалить из выборки ячейки с буквой М, т.е. те ячейки, в которых нет выборочных значений.

Недостатком данного метода генерирования случайных чисел является то, что невозможно заранее предсказать, сколько значений будет в конечной выборке. Существуют различные модификации этого метода, уменьшающие количество пробных пар равномерно распределенных случайных чисел.

	C2		=ЕСЛИ(B2^2<1-(2*A2-1)^2,2*A2-1,"M")		
	A	B	C	D	E
	Первое случайное число	Второе случайное число	Выборка		
1					
2	0,74337354	0,38423094	0,486747		
3	0,98245872	0,76354456	M		
4	0,77353041	0,93845456	M		
5	0,10498207	0,05344605	-0,79004		
6	0,89769734	0,24026946	0,795395		
7	0,69190828	0,49694334	0,383817		
8	0,22030094	0,44252204	-0,5594		
9	0,2684171	0,78689106	-0,46317		
10	0,33812618	0,33562028	-0,32375		
11	0,61814882	0,94851918	0,236298		
12	0,19128499	0,05090927	-0,61743		
13	0,36551464	0,82121517	-0,26897		
14	0,79301941	0,9306162	M		
15	0,85613071	0,45490732	0,712261		
16	0,23543151	0,27105428	-0,52914		

Рис. 7.15. Генерирование случайных чисел по методу отбора

7.5. Моделирование многомерных случайных величин

Если компоненты X_1, X_2, \dots, X_n многомерной случайной величины $X = (X_1, X_2, \dots, X_n)$ независимы, то можно моделировать каждую случайную величину X_k ($k = 1, 2, \dots, n$) независимо и из реализаций этих величин (сгенерированных случайных чисел) составить ряд n -мерных векторов, которые образуют выборку из генеральной совокупности случайной величины X .

В случае зависимости компонентов X_1, X_2, \dots, X_n для моделирования многомерной случайной величины $X = (X_1, X_2, \dots, X_n)$ необходимо использовать совместную функцию распределения $F(x_1, x_2, \dots, x_n)$ компонентов X_1, X_2, \dots, X_n . Для упрощения выкладок далее рассмотрим случай двумерной непрерывной случайной величины $X = (X_1, X_2)$; моделирование в общем случае показано в [16, глава 2].

Пусть совместная функция распределения $F(x_1, x_2)$ дважды дифференцирована и существует совместная плотность вероятности $f(x_1, x_2) = \frac{\partial^2 F(x_1, x_2)}{\partial x_1 \partial x_2}$. Эту

плотность можно представить в виде произведения частной и условной плотностей вероятностей случайных величин X_1 и X_2 :

$$f(x_1, x_2) = f_1(x_1) f_2(x_2 | x_1) = f_2(x_2) f_1(x_1 | x_2),$$

где $f_1(x_1) = \int_{-\infty}^{\infty} f(x_1, x_2) dx_2$, $f_2(x_2 | x_1) = \frac{f(x_1, x_2)}{f_1(x_1)}$ (функции $f_2(x_2)$ и $f_1(x_1 | x_2)$ вычис-

ляются по аналогичным формулам с заменой индексов 1 на 2 и 2 на 1). Далее необходимо вычислить условные функции распределения:

$$F_1(x) = \int_{-\infty}^x f_1(x_1) dx_1, \quad F_2(x | x_1) = \int_{-\infty}^x f_2(x_2 | x_1) dx_2.$$

Моделирование случайной величины $X = (X_1, X_2)$ основано на том факте, что случайные величины X_1 и X_2 , полученные при последовательном решении уравнений $F_1(X_1) = Y_1$, $F_2(X_2 | X_1) = Y_2$, где Y_1 и Y_2 — независимые равномерно распределенные на интервале $[0, 1]$ случайные величины, имеют совместную функцию распределения $F(x_1, x_2)$.

Рассмотрим пример моделирования случайной величины $X = (X_1, X_2)$, которая может принимать значения в треугольнике $x + y = 1$, $x > 0$, $y > 0$ с плотностью вероятности $f(x, y) = 6y$. Сначала вычислим условные плотности вероятностей:

$$f_1(x) = \int_{-\infty}^{+\infty} f(x, y) dy = \int_0^{1-x} 6y dy = 3(1-x)^2, \quad f_2(y | x) = \frac{f(x, y)}{f_1(x)} = \frac{2y}{(1-x)^2}.$$

Далее вычислим условные функции распределения:

$$F_1(x) = \int_{-\infty}^x f_1(u) du = \int_0^x 3(1-u)^2 du = 1 - (1-x)^3,$$

$$F_2(y | x) = \int_{-\infty}^y f_2(u | x) du = \frac{2}{(1-x)^2} \int_0^y u du = \frac{y^2}{(1-x)^2}.$$

В данном случае решение уравнений $F_1(x) = \xi_1$, $F_2(y | x) = \xi_2$ можно найти в явном виде: $x = 1 - \sqrt[3]{1 - \xi_1}$, $y = \sqrt[3]{1 - \xi_1} \sqrt{\xi_2}$. Поскольку случайные величины ξ_1 и $1 - \xi_1$ имеют одинаковые распределения, если величина ξ_1 равномерно распределена на интервале $[0, 1]$, формулы для вычисления x и y можно записать следующим образом: $x = 1 - \sqrt[3]{\xi_1}$, $y = \sqrt[3]{\xi_1} \sqrt{\xi_2}$.

На рис. 7.16 показан рабочий лист, в котором сгенерированы 15 значений случайной величины $X = (X_1, X_2)$. Значения случайных величин X_1 и X_2 получены с помощью формул массива

$$\{=1-\text{СТЕПЕНЬ}(\text{СЛЧИС}();1/3)\} \text{ и } \{=(1-\text{A2}:\text{A16})*\text{КОРЕНЬ}(\text{СЛЧИС}())\}$$

соответственно.

Отметим, что для решения уравнений $F_1(X_1) = Y_1$ и $F_2(X_2 | X_1) = Y_2$ в общем случае в Excel можно использовать средство Поиск решения.

7.5.1. Моделирование зависимых случайных величин с известным коэффициентом корреляции

Описанный ниже метод обычно применяется для моделирования зависимых случайных величин, распределение которых принадлежит классу *безгранично делимых распределений*, и реже — для случайных величин с другими распределениями, поскольку в последнем случае, как правило, необходимы предварительные достаточно сложные аналитические выкладки. Если распределение F случайной величины X принадлежит классу безгранично делимых распределений, то случайную величину X можно представить как сумму независимых одинаково распределенных случайных величин $X = X_1 + X_2$, имеющих тот же тип распределения F (возможно, с другими параметрами). Справедливо и обратное утверждение: если случайные величины X_1 и X_2 имеют один и тот же тип распределения F , принадлежащий классу безгранично делимых распределений, то случайная величина $X = X_1 + X_2$ имеет тот же тип распределения F . Классу

безгранично делимых распределений принадлежат многие распределения, встречающиеся на практике, в частности, нормальное распределение, распределение Пуассона, биномиальное распределение, гамма-распределение, распределение χ^2 и др.

B2		A:=(1-A2:A16)*КОРЕНЬ(СЛЧИС())				
	A	B	C	D	E	F
1	X1	X2				
2	0,410571	0,372824				
3	0,164295	0,664175				
4	0,311271	0,343393				
5	0,113769	0,879303				
6	0,490428	0,372845				
7	0,111392	0,755352				
8	0,346868	0,326384				
9	0,217042	0,531696				
10	0,861867	0,87895				
11	0,54171	0,30385				
12	0,191951	0,689284				
13	0,046275	0,029688				
14	0,024201	0,723018				
15	0,336923	0,609585				
16	0,270098	0,184518				
17						

Рис. 7.16. Моделирование случайной величины $X = (X_1, X_2)$

Пусть случайные величины X_1, X_2, \dots, X_n , являющиеся компонентами многомерной случайной величины $X = (X_1, X_2, \dots, X_n)$, имеют математические ожидания $m = (m_1, m_2, \dots, m_n)$ и среднеквадратические отклонения $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_n)$. Их совместное распределение характеризуется корреляционной матрицей $R = \{r_{ij}\}$ ($i, j = 1, 2, \dots, n$), где $r_{ii} = 1$, если $i = j$, а при $i \neq j$ r_{ij} они являются коэффициентами корреляции между случайными величинами X_i и X_j . Известно, что матрицу R можно представить в виде произведения двух треугольных матриц: $R = SS^T$. Обозначим как $Y = (Y_1, Y_2, \dots, Y_n)$ вектор независимых случайных величин, имеющих нулевые математические ожидания и единичные дисперсии. Тогда случайная величина $Z = \sigma SY + m$ будет иметь векторы математических ожиданий $m = (m_1, m_2, \dots, m_n)$ и среднеквадратических отклонений $\sigma = (\sigma_1, \sigma_2, \dots, \sigma_n)$, а зависимость между компонентами этого вектора будет характеризоваться корреляционной матрицей R .

Чтобы на основе преобразования $Z = \sigma SY + m$ полностью смоделировать распределение случайной величины X , необходимо так подобрать распределения случайных величин $Y = (Y_1, Y_2, \dots, Y_n)$, чтобы частные распределения компонентов Z_1, Z_2, \dots, Z_n , составляющих вектор Z , совпадали с частными распределениями величин X_1, X_2, \dots, X_n . Условие принадлежности этих распределений одному типу распределений из класса безгранично делимых законов значительно облегчает решение такой задачи. Обозначим как $\bar{X}_1, \bar{X}_2, \dots, \bar{X}_n$ стандартизованные

Матрица R является положительно определенной матрицей, поэтому ее можно представить в виде такого произведения.

случайные величины, имеющие нулевые математические ожидания, единичные дисперсии и такие же распределения, как и величины X_1, X_2, \dots, X_n .

Пусть матрица S — нижняя треугольная; тогда все ее элементы, стоящие выше главной диагонали, равны нулю, и, поскольку эта матрица является разложением корреляционной матрицы, сумма квадратов элементов любой строки матрицы равна 1. Тогда

$$SY = (Y_1, s_{21}Y_1 + s_{22}Y_2, s_{31}Y_1 + s_{32}Y_2 + s_{33}Y_3, \dots, s_{n1}Y_1 + s_{n2}Y_2 + \dots + s_{nn}Y_n).$$

Получаем, что

$$Y_1 = \bar{X}_1, Y_2 = (\bar{X}_2 - s_{21}Y_1)/s_{22}, Y_3 = (\bar{X}_3 - s_{31}Y_1 - s_{32}Y_2)/s_{33}, \dots,$$

$$Y_n = (\bar{X}_n - s_{n1}Y_1 - s_{n2}Y_2 - \dots - s_{n(n-1)}Y_{n-1}).$$

Отсюда видно, что распределение случайной величины Y_1 совпадает с распределением случайной величины \bar{X}_1 . Зная распределение случайных величин Y_1 и \bar{X}_2 , можно найти распределение случайной величины Y_2 и т.д. Таким образом можно последовательно найти распределения всех величин Y_1, Y_2, \dots, Y_n . Эти распределения можно найти для любых распределений величин X_1, X_2, \dots, X_n , хотя, может быть, с некоторыми сложностями. Но еще раз подчеркнем, что наиболее просто эти распределения определяются в случае, когда распределения величин X_1, X_2, \dots, X_n принадлежат одному типу безгранично делимых распределений. В этом случае достаточно вычислить только параметры распределений, а не определять тип распределений.

Покажем реализацию описанного метода в упрощенном (но наиболее часто используемом на практике) варианте, когда $n = 2$, а случайная величина $X = (X_1, X_2)$ имеет двумерное нормальное распределение. Пусть случайные величины X_1 и X_2 имеют математические ожидания и среднеквадратические отклонения соответственно m_1, m_2 и σ_1, σ_2 . Коэффициент корреляции между ними

пусть равен r . Тогда матрица S имеет вид $S = \begin{pmatrix} 1 & 0 \\ r & \sqrt{1-r^2} \end{pmatrix}$.

Случайные величины Y_1 и Y_2 в данном случае будут иметь стандартные нормальные распределения. Значения случайной величины X_1 будут вычисляться по формуле $x_1 = \sigma_1 y_1 + m_1$, а величины X_2 — по формуле $x_2 = \sigma_2 (r y_1 + \sqrt{1-r^2} y_2) + m_2$, где y_1 и y_2 — значения случайных величин Y_1 и Y_2 . На рис. 7.17 последовательно показаны вычисления случайных чисел x_1 и x_2 . В столбце $sY1$ по формуле массива $\{=НОРМСТОБР(СЛЧИС())\}$ вычисляются значения случайной величины Y_1 (диапазон ячеек, содержащий эти значения, назван $sY1$). В столбце $sY2$ с использованием формулы массива

$$\{=F1*sY1+КОРЕНЬ(1-F1*F1)*НОРМСТОБР(СЛЧИС())\}$$

вычисляются значения $r y_1 + \sqrt{1-r^2} y_2$ (значение коэффициента корреляции записано в ячейке $F1$). Диапазон ячеек, содержащий эти значения, назван $sY2$. В столбцах $X1$ и $X2$ вычисляются значения величин X_1 и X_2 по формулам массивов $\{=F4*sY1+F2\}$ и $\{=F5*sY2+F3\}$ соответственно. Конечно, можно обойтись без промежуточных вычислений в столбцах A и B , однако такие вычисления более

наглядны и просты. Кроме того, они позволяют легко генерировать выборки с различными параметрами распределения, для чего достаточно изменить значения в ячейках F1:F5. В ячейке F8 вычисляется выборочный коэффициент корреляции. Как видно на рис. 7.17, подсчитанное значение выборочного коэффициента корреляции близко к истинному значению коэффициента корреляции.

F8		=КОРРЕЛ(X1,X2)					
	A	B	C	D	E	F	G
1	sY1	sY2	X1	X2	r =	-0,5	
2	0,226589	1,311807	0,453178	2,311807	m1 =	0	
3	2,533087	-0,83322	5,066174	0,166779	m2 =	1	
4	0,9502	0,158277	1,9004	1,158277	σ1 =	2	
5	-0,06775	-0,06823	-0,13551	0,931769	σ2 =	1	
6	-1,49732	1,319672	-2,99463	2,319672	Выборочный коэффициент корреляции		
7	0,409742	-0,722	0,819485	0,278003			
8	-0,52151	0,048427	-1,04301	1,048427		-0,49571	
9	0,827951	-1,11062	1,655903	-0,11062			
10	-0,07023	0,639052	-0,14047	1,639052			
11	0,078091	-0,34319	0,156182	0,656805			
12	-1,14213	0,626988	-2,28427	1,626988			
13	-1,27624	0,645726	-2,55247	1,645726			
14	0,373679	-0,36189	0,747359	0,638105			
15	1,302126	-1,7303	2,604252	-0,7303			
16	0,814497	-0,22235	1,628994	0,77765			
17	-2,49346	0,869561	-4,98692	1,869561			

Рис. 7.17. Моделирование двумерного нормального распределения

Анализ одномерных выборок

В этой части...

Глава 8. Предварительный анализ

Глава 9. Подбор распределения

Глава 10. Интервальное оценивание параметров распределения

Глава 11. Проверка гипотез о параметрах распределений

Глава 12. Сравнение одномерных выборок

В этой части речь идет о практической реализации методов статистического анализа одномерных независимых выборок. Глава 8 посвящена предварительной обработке данных, в главе 9 рассмотрены важные для последующего анализа вопросы подбора распределений по имеющимся выборочным значениям. В главах 10 и 11 показаны методы интервального оценивания параметров распределений и критерии проверки гипотез о значениях этих параметров. Глава 12 посвящена сравнению распределений нескольких одномерных выборок.

Предварительный анализ

Предварительный этап на практике присутствует в любом статистическом анализе и заключается, по крайней мере, в подготовке данных для проведения анализа. Но, конечно, действия, выполняемые на этом этапе, зависят от конкретных задач, решаемых с помощью статистического анализа, и исходных данных. Так, если возникли подозрения, что выборка имеет значения, которые не являются случайными или резко выделяются на фоне остальных выборочных значений, то следует провести цензурирование выборки. Если необходима интервальная оценка неизвестных параметров распределения, то предварительным этапом можно считать проверку гипотезы о нормальности выборочного распределения, чтобы впоследствии использовать интервальные оценки, построенные на основе нормального распределения. В других случаях целью статистического анализа может быть установка типа выборочного распределения, а на предварительном этапе строятся гистограммы выборочного распределения и подсчитываются различные статистические характеристики выборки, чтобы подобрать тип распределения, наиболее соответствующий исходным данным.

Таким образом, с одной стороны, трудно очертить круг тех действий, которые следует *всегда* выполнять в качестве предварительного анализа; с другой стороны, невозможно четко отделить предварительный этап статистического анализа от самого анализа. Поэтому здесь мы остановимся лишь на некоторых действиях, которые обычно относят к предварительному этапу статистического анализа. Рассмотрим цензурирование и преобразование выборок, построение гистограмм, полигонов и эмпирических функций распределения, а также вычисление точечных оценок параметров выборочных распределений. Другими словами, рассмотрим "техническую" работу, проводимую перед применением статистических методов.

8.1. Цензурирование

Иногда в данных можно наблюдать *выбросы* — сильно отклоняющиеся значения, т.е. значения, которые, по-видимому, не принадлежат данному распределению, поскольку они либо слишком велики, либо слишком малы. Выбросы затрудняют проводимый статистический анализ и могут привести к неверно интерпретируемым результатам. Поэтому выбросы следует выявить и обработать отдельно. Процесс удаления из выборки выбросов называется *цензурированием выборки*.

В зависимости от предположений о природе выбросов (это ошибки наблюдений или артефакты, привнесенные человеком, либо корректные, но "отличающиеся от остальных" значения данных) проблему выбросов решают по-разному. Но в любом случае предпринимаемые действия по решению этой

проблемы *необходимо обосновывать* исходя либо из природы выбросов, либо из целей конкретного статистического анализа.

Если это элементарная ошибка наблюдений, то значение по возможности нужно просто откорректировать. Если это артефакт, не подлежащий корректровке, то его удаляют. Если есть убедительные подтверждения тому, что значения-выбросы не принадлежат генеральной совокупности, из которой получена исследуемая выборка, то их также удаляют. Если последнее утверждение обосновать трудно, но все-таки есть "подозрительные" выборочные значения, то можно выполнить два анализа — без удаления выбросов и с удалением выбросов.

Следует отметить, что для цензурированных выборок иногда применяют специальные формулы для вычисления оценок параметров распределения [14]. Часто эти вычисления выполняются итерационно, пока не сойдутся к определенным значениям. Применение таких формул обычно требует априорных предположений о типе распределения; "универсальные" формулы весьма сложны [23] и на практике используются редко. Такие формулы мы рассматривать не будем.

8.1.1. Цензурирования на основе построения доверительных интервалов

Существует несколько основных подходов к идентификации выбросов, среди которых выделим подход, основанный на априорной информации о распределении генеральной совокупности, и непараметрические методы, не использующие информации о распределении генеральной совокупности. Рассмотрим сначала первый подход при самых общих предположениях.

Идея выделения выбросов среди выборочных значений достаточно проста. На основе выборки каким-либо образом строится доверительный интервал, содержащий основную массу значений с заданной вероятностью. Значения, выходящие за этот интервал, считаются выбросами. Затем на основании уже цензурированной выборки строится новый доверительный интервал и выборка снова проверяется на наличие выбросов. Если таковые имеются, то процесс повторяется до тех пор, пока объем цензурированной выборки не стабилизируется, т.е. до тех пор, пока будут идентифицироваться новые выбросы. В этом методе очевидна роль априорных предположений о типе распределения генеральной совокупности, поскольку на основе этих предположений строится доверительный интервал. Еще одной проблемой является неизвестность значений параметров распределения, вместо которых приходится брать их выборочные оценки. Это, в свою очередь, приводит к требованию достаточно большого объема выборки. Отметим также, что вместо стандартной оценки среднеквадратического отклонения как корня из выборочной дисперсии рекомендуется использовать среднее абсолютное отклонение

$$d_n = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|, \text{ где } \bar{x} — \text{выборочное среднее [1, 23], особенно для малых выбо-}$$

рок и выборок, распределение которых значительно отличается от нормального.

Если не делать ограничительных предположений о типе распределения, то единственным способом построения доверительного интервала является использование неравенства Чебышева или неравенства Гаусса в предположении, что распределение одномодально (эти неравенства приведены в разделе 1.2.4). Чтобы составить представление о виде распределения, перед началом цензурирования следует построить гистограмму, которая по крайней мере покажет, можно ли

считать распределение одномодальным. Отметим, что использование неравенства Чебышева — наиболее надежный и безопасный способ цензурирования, поскольку в этом случае вероятность отбросить те значения, которые действительно принадлежат выборке, минимальна (но, с другой стороны, максимальна вероятность оставить выбросы в выборке).

На рис. 8.1 в столбце А показана выборка объемом 50 значений, имеющая логнормальное распределение с параметрами $m = 0$ и $\sigma = 1$, к которой добавлены значения -0,5, -1,2, 8, 9 и 10. (Выборочные значения получены в результате применения формулы массива $\{=\text{ЛОГНОРБР}(\text{СЛЧИС}();0;)\}$ к диапазону А2:А51. Затем формулы были заменены значениями, как описано в разделе 7.1.) Отметим, что здесь отрицательные значения являются очевидным артефактом, однако только если априори известно, что выборка является реализацией случайной величины, принимающей положительные значения с вероятностью 1. В противном случае исключить отрицательные значения из выборки "законным" способом практически невозможно. В столбце В на рис. 8.1 приведена та же выборка, отсортированная в порядке возрастания; результаты цензурирования не зависят от порядка выборочных значений, но для наглядности удобнее использовать отсортированную выборку.

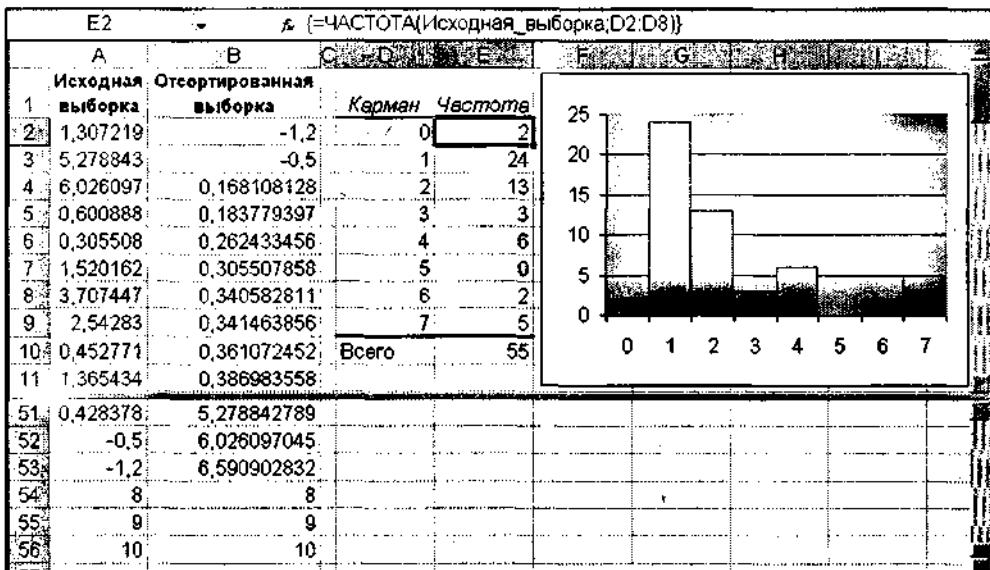


Рис. 8.1. Исходные данные для цензурирования

На этом же рисунке показана гистограмма частот. (Для подсчета частот в диапазоне Е2:Е9 использована формула массива $\{=\text{ЧАСТОТА}(\text{Исходная_выборка};0;2;0;8)\}$; диапазон ячеек, содержащий выборочные значения, назван Исходная_выборка.) Значения, превышающие 6, сгруппированы в интервале, помеченном числом 7. На гистограмме видно, что распределение далеко от одномодального, поэтому используем неравенство Чебышева.

Напомним, что неравенство Чебышева (см. раздел 1.2.4) имеет вид

$$P(|X - m| \geq \lambda \sigma) \leq 1/\lambda^2,$$

где X — случайная величина, m — ее математическое ожидание, a — средне-квадратическое отклонение, X определяет размер доверительного интервала и вычисляется на основании заданного доверительного уровня (вероятности) p . В качестве оценки математического ожидания используем выборочное среднее, а вместо среднеквадратического отклонения — среднее абсолютное отклонение d_n . Если задана вероятность p , с которой доверительный интервал должен содержать основную массу выборочных значений, то далее значение $1 - p$ приравнивается к $1/\Pi^2$ и из этого равенства определяется значение A . Таким образом, A вычисляется по формуле $X = l/\sqrt{1-p}$, нижняя t_n и верхняя t_b границы доверительного интервала вычисляются по формулам

$$t_n = \bar{x} - Xd_n \text{ и } t_b = \bar{x} + Xd_n.$$

Рабочий лист с результатами вычислений по этим формулам для $p = 0,9$ показан на рис. 8.2. Теперь осталось определить, какие выборочные значения выходят за построенный доверительный интервал. Конечно, для данной относительно малой выборки это сделать несложно, тем более что она отсортирована. И все-таки покажем два способа автоматизации процесса поиска выбросов.

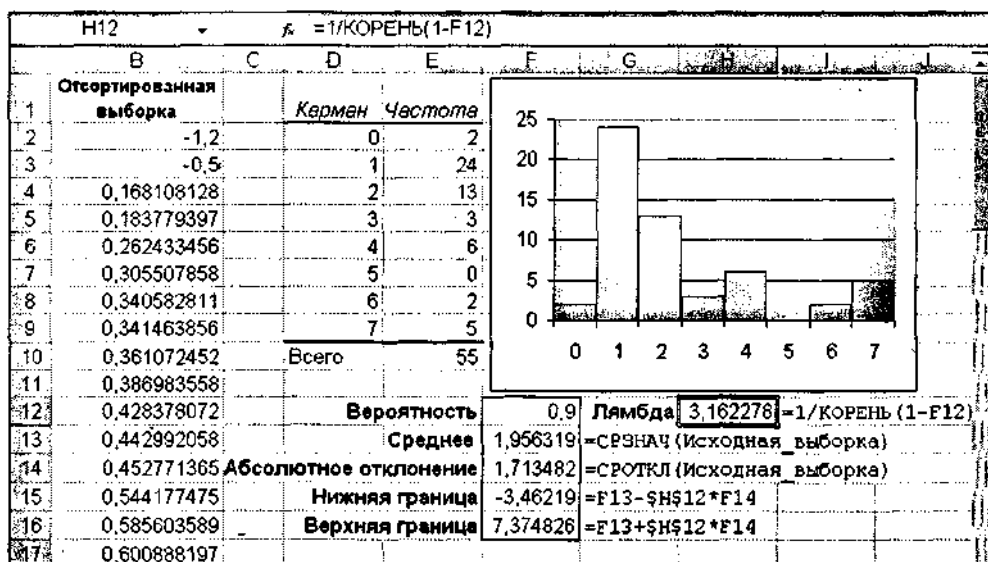


Рис. 8.2. Результаты вычислений

Первый способ просто выделяет на основе заранее заданного формата значения, выходящие за границы доверительного интервала. Для этого используется условное форматирование.

1. Сначала необходимо выделить диапазоны ячеек, к которым будет применено условное форматирование; в данном примере это диапазон A1:B56, содержащий как исходную, так и отсортированную выборки.
2. По команде **Формат** → **Условное форматирование** открывается одноименное диалоговое окно (рис. 8.3). В нем необходимо задать условие, которому должны

удовлетворять значения, чтобы к этим значениям был применен определенный формат, и сам формат. Для задания условия в первом поле следует указать, что условие задается относительно значения (можно также задать условие в виде формулы), во втором поле из раскрывающегося списка необходимо выбрать знак равенства или неравенства. В третьем поле нужно указать значение, с которым сравнивается значение в ячейке. Здесь можно ввести не только конкретное число, но и ссылку на ячейку, содержащую это число. В данном примере первое условие задается для значений, которые меньше нижней границы; само значение нижней границы вычислено в ячейке F15.

3. Чтобы задать формат, надо щелкнуть на кнопке **Формат**, после чего откроется диалоговое окно **Формат ячеек**. В нем можно задать любой формат как для значений, так и для ячеек, их содержащих.
4. Для задания еще одного условия и соответствующего формата (например, чтобы по-разному форматировать наибольшие и наименьшие значения) следует щелкнуть на кнопке **А также >>**. Окно расширится, и можно будет задать новые условие и формат. В данном примере второе условие задается для значений, которые больше верхней границы (ячейка F16).
5. После задания всех условий и форматов следует щелкнуть на кнопке **ОК**. Форматы будут немедленно применены к выделенному диапазону ячеек. Результат применения условного формата для описываемого примера показан на рис. 8.4.

Всего можно задать до трех условий. Ячейки, содержимое которых не удовлетворяет ни одному условию, сохраняют формат, который они имели до задания условного формата. Достоинством условного форматирования является то, что при изменении значений в ячейках, содержащих как выборочные, так и вычисляемые значения (например, значения нижней и верхней границ), условное форматирование сохраняется и применяется к новым значениям.

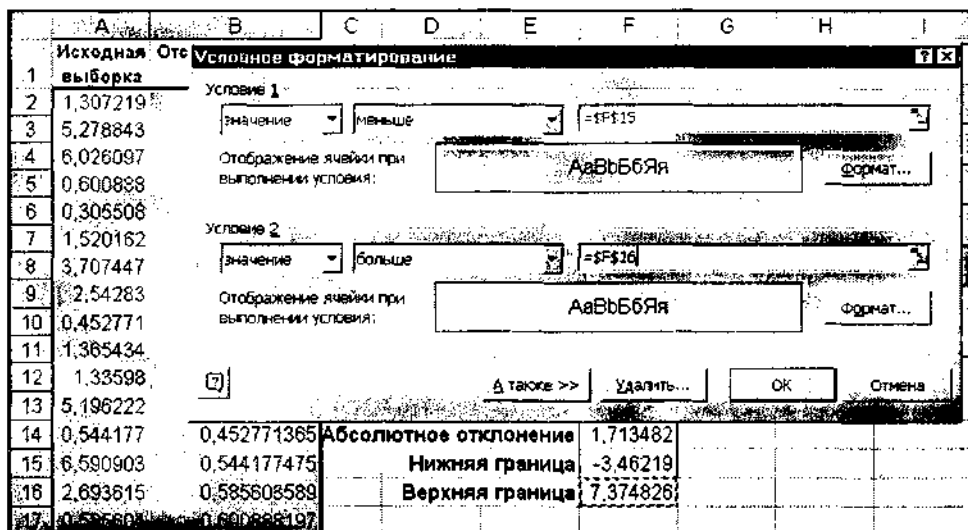


Рис. 8.3. Задание условных форматов

	A	B	C	D	E	F	G	H
7	1,520162	0,305507858		5	0	5		
8	3,707447	0,340582811		6	2	0		
9	2,54283	0,341463856		7	5	0		
10	0,452771	0,361072452	Всего	55				
11	1,365434	0,386983558						
12	1,33598	0,428378072		Вероятность	0,9	Лямбда	3,162278	
13	5,196222	0,442992058		Среднее	1,956319			
14	0,544177	0,452771365	Абсолютное отклонение		1,713482			
15	6,590903	0,544177475	Нижняя граница		-3,46219			
16	2,693615	0,585603589	Верхняя граница		7,374826			
17	0,585604	0,600888197						
51	0,428378	5,278842789						
52	-0,5	6,026097045						
53	-1,2	6,590902832						
54	8	8						
55	9	9						
56	10	10						

Рис. 8.4. Применение условного форматирования

Второй способ исключения выбросов более радикален, поскольку он формирует новую выборку, но уже без этих выбросов. Чтобы создать такую выборку, выполните следующие действия.

1. Выделите диапазон ячеек, совпадающий по размеру с диапазоном, содержащим исходную выборку.
2. Введите приведенную ниже формулу и нажмите клавиши <Ctrl+Shift+Enter>. Тем самым будет создана формула массива, распространяющая свое действие на весь выделенный диапазон. (Здесь диапазон ячеек, содержащий выборку, назван Выборка.)

=ЕСЛИ(Выборка<P15;"";ЕСЛИ(Выборка>P16;"";Выборка))

Как показано на рис. 8.5, эта формула оставляет ячейки, которые соответствуют значениям, выходящим за доверительный интервал, пустыми. (На рис. 8.5 также показана гистограмма для новой выборки.) После принятия окончательного решения о том, чтобы оставить цензурированную выборку (принятое, конечно, после Дополнительных экспериментов со значением вероятности p), можно удалить из этого диапазона формулы и оставить только значения.

Если задать значение вероятности p равным 0,95 (что более естественно, чем значение 0,9), то в этом случае будет исключено только значение 10, как показано на рис. 8.6. Таким образом, неравенство Чебышева слишком "осторожно" в определении выбросов.

Для построения доверительного интервала, содержащего основную массу выборочных значений, можно также применить эмпирическое правило $3S$, которое утверждает, что вероятность $P\{|X - \bar{x}| < 3S_n\}$ составляет не менее 0,95. Здесь \bar{x} — выборочное среднее, а S_n — выборочная оценка среднеквадратического отклонения. Результаты цензурирования на основе этого неравенства показаны на рис. 8.7. В данном случае цензурированная выборка совпадает с выборкой,

полученной при использовании неравенства Чебышева с вероятностью $p = 0,9$. Но следует отметить, что вместо выборочной оценки среднеквадратического отклонения здесь по-прежнему использовалось среднее абсолютное отклонение d_n .

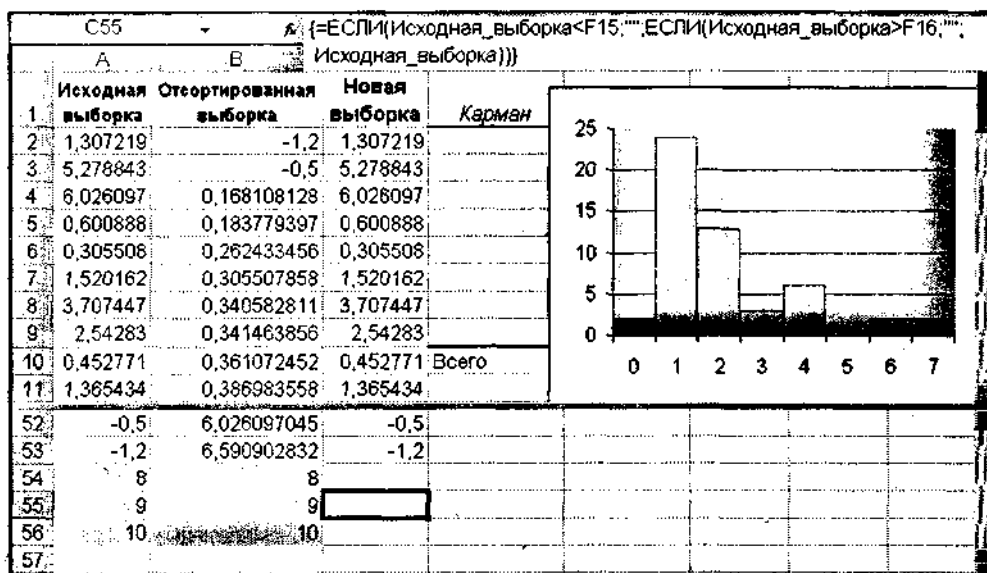


Рис. 8.5. Новая цензурированная выборка

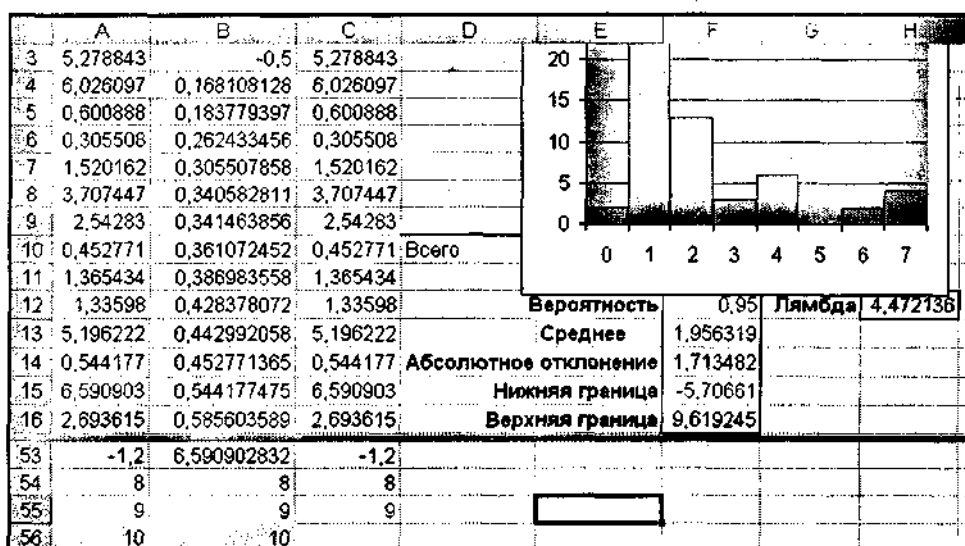


Рис. 8.6. Цензурированная выборка при вероятности 0,95

В заключение отметим, что при проведении цензурирования на основе неравенств Чебышева, Гаусса или на основе правила 3S итерационная процедура цензурирования (последовательного проведения цензурирования до тех пор, пока будут определяться новые выбросы), как правило, не проводится, поскольку

здесь не предусмотрены "стабилизирующие" поправки для вычисления среднего и оценки среднеквадратического отклонения, аналогичные применяемым при цензурировании выборок из нормально распределенных совокупностей [23].

F14		=F12-3*F13			
	A	B	C	D	E
5	0,600888	0,183779397	0,600888	3	3
6	0,305508	0,262433456	0,305508	4	6
7	1,520162	0,305507858	1,520162	5	0
8	3,707447	0,340582811	3,707447	6	2
9	2,54283	0,341463856	2,54283	7	2
10	0,452771	0,361072452	0,452771	Всего	52
11	1,365434	0,386983558	1,365434		
12	1,33598	0,428378072	1,33598	Среднее	1,956319
13	5,196222	0,442992058	5,196222	Среднее отклонение	1,713482
14	0,544177	0,452771365	0,544177	Нижняя граница	-3,18413
15	6,590903	0,544177475	6,590903	Верхняя граница	7,096766
16	2,693615	0,585603589	2,693615		
52	-0,5	6,026097045	-0,5		
53	-1,2	6,590902832	-1,2		
54	8	8			
55	9	9			
56	10	10			
57					

Рис. 8.7. Цензурирование на основе правила 3S

8.1.2. Непараметрическое цензурирование

В описанном ниже методе цензурирования не требуется априорных предположений о распределении генеральной совокупности, поскольку он построен на основе порядковых статистик (о порядковых статистиках речь идет в разделе 2.3.9). Существует несколько подходов к цензурированию выборок на основе порядковых статистик. Покажем метод цензурирования, предложенный Дж. Тьюки (J.W. Tukey) [15]. Для иллюстрации метода используем ту же выборку, что и в предыдущем разделе.

1. Вычисляется ранг $r_{0,25}$ нижнего квартиля $\xi_{0,25}$ (о квартилях речь идет в разделе 1.2.3) по формуле $r_{0,25} = (1 + [(1 + n)/2])/2$, где n — объем выборки, $[x]$ обозначает целую часть числа x . На рис. 8.8 этот ранг вычисляется в ячейке E2 по формуле $= (1 + \text{ЦЕЛОЕ}((1 + \$D\$2)/2))/2$ (в ячейке D2 подсчитывается объем выборки по формуле $= \text{СЧЁТ}(\text{Исходная_выборка})$).
2. Вычисляется ранг $r_{0,75}$ верхнего квартиля $\xi_{0,75}$ по формуле $r_{0,75} = n + 1 - r_{0,25}$. На рабочем листе, показанном на рис. 8.8, данный ранг вычисляется в ячейке F2 по формуле $= \$D\$2 + 1 - E2$.
3. Определяются значения нижнего квартиля $\xi_{0,25}$ и верхнего квартиля $\xi_{0,75}$: если вычисленные ранги этих квартилей — целые числа, то в качестве значений этих квартилей берутся выборочные значения рангов, совпадающих с вычисленными рангами квартилей. Если же вычисленные ранги квартилей дробные, то в качестве значений квартилей берется среднее выборочных значений с рангами, ближайших к вычисленным рангам квартилей. Например, если ранг нижнего квартиля равен 14,5 (как в нашем

примере), за значение этого квартиля принимается среднее выборочных значений с рангами 14 и 15. Чтобы автоматизировать определение значений квартилей и реализовать эти простые вычисления, в электронной таблице приходится применять достаточно сложные формулы. На рис. 8.8 в ячейке E4 для вычисления нижнего квартиля используется формула

=ЕСЛИ(E2-ЦЕЛОЕ(E2)=0;ИНДЕКС(B2:B56;E2;1);
(ИНДЕКС(B2:B56;ЦЕЛОЕ(E2);1)+ИНДЕКС(B2:B56;ЦЕЛОЕ(E2)+1;1))/2).

Аналогичная формула используется в ячейке F4 для вычисления верхнего квартиля:

=ЕСЛИ(F2-ЦЕЛОЕ(F2)=0;ИНДЕКС(B2:B56;F2;1);
(ИНДЕКС(B2:B56;ЦЕЛОЕ(F2);1)+ИНДЕКС(B2:B56;ЦЕЛОЕ(F2)+1;1))/2).

(В формулах применена функция ИНДЕКС в форме массива. Эта функция в данной форме возвращает содержимое ячейки, расположенной на пересечении указанной строки и указанного столбца (второй и третий аргументы функции) диапазона ячеек, задаваемого в первом аргументе функции.)

4. Вычисляются нижняя t_n и верхняя t_n границы, относительно которых определяются выборочные значения, принимаемые за выбросы; выбросами считаются значения, которые меньше t_n и которые больше t_n . Эти границы вычисляются по формулам¹

$$t_n = \xi_{0,25} - 1,5(\xi_{0,25} - \xi_{0,75}) \text{ и } t_n = \xi_{0,75} + 1,5(\xi_{0,25} - \xi_{0,75}).$$

В нашем примере данные значения вычисляются соответственно в ячейках E6 и F6 по формулам =E4-1,5*(F\$4-\$E\$4) и =F4+1,5*(F\$4-\$E\$4).

5. В исходной выборке вычисляются значения, которые выходят за нижнюю и верхнюю границы. Это можно сделать способами, описанными в предыдущем разделе. На рис. 8.9 показаны выбросы, выделенные с помощью условного форматирования.

Как видно на рис. 8.9, в результате цензурирования в качестве выбросов определены и два “правильных” выборочных значения. Это еще раз доказывает, что к цензурированию надо относиться осторожно и применять его следует только тогда, когда для этого есть веские основания.

8.1.3. Винзоризация выборки

Винзоризация выборки, являясь своеобразной разновидностью цензурирования, отличается от последней тем, что идентифицированные выбросы не удаляются из выборки; им присваиваются значения, равные соответственно нижней t_n или верхней t_n границам, относительно которых идентифицируются выбросы. Часто винзоризацию выполняют при задании только одной границы — верхней или нижней. Это называется односторонней винзоризацией, в отличие от двухсторонней, когда используются обе границы.

При задании границ применяются два подхода. При первом подходе исходя из каких-либо априорных предположений задается количество выборочных значений, которые будут винзоризируемы (т.е. приравнены к значениям границ). Например, можно задать, что винзоризируется не более 5% или 10% выборочных

¹ Разность между квартилями $\xi_{0,75} - \xi_{0,25}$ называется интерквартильный размах и иногда используется в качестве меры разброса выборочных значений.

значений. Такой способ задания границ реализован на рабочем листе, показанном на рис. 8.10 и 8.11. (На этих рисунках показаны и расчетные формулы.) Отметим, что здесь вычисляются не сами значения границ, а нижний и верхний ранги выборочных значений, которые определяют границы. Такой подход немного упрощает вычисления.

F6 =F4+1,5*(\$F\$4-\$E\$4)						
	A	B	C	D	E	F
1	Исходная выборка	Отсортированная выборка	Ранги	Объем	Ранг нижнего квартиля	Ранг верхнего квартиля
2	1.307219		-1,2	1	55	41,5
3	5.278843		-0,5	2	Нижний квартиль	Верхний квартиль
4	6.026097	0,168108128	3		0,564890532	2,618222482
5	0,600888	0,183779397	4		Нижняя граница	Верхняя граница
6	0,305508	0,262433456	5		-2,515107392	5,698220406
7	1,520162	0,305507858	6			
8	3,707447	0,340582811	7			
9	2,54283	0,341463856	8			
10	0,452771	0,361072452	9			
11	1,365434	0,386983558	10			
12	1,33598	0,428378072	11			
13	5,196222	0,442992058	12			
14	0,544177	0,452771365	13			
15	6,590903	0,544177475	14			
16	2,693615	0,585603589	15			

Рис. 8.8. Подготовка к цензурированию

Другой подход к определению нижней t_n и верхней t_v границ аналогичен построению доверительного интервала на основе выборочных оценок математического ожидания и дисперсии (тогда t_n и t_v будут границами данного интервала). На этой основе организована следующая итерационная процедура винзоризации выборочных значений [23].

1. По выборке вычисляются выборочное среднее $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ и выборочное среднеквадратическое отклонение s_n как корень из выборочной дисперсии

$$s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2 \quad (x_i - \text{выборочные значения, } n - \text{объем выборки}).$$

Также вычисляются границы $t_n = \bar{x} - cs_n$ и $t_v = \bar{x} + cs_n$, где константа c определяет сходимость процедуры (поскольку это итерационная процедура) и обычно выбирается из интервала от 1 до 2, например 1,5.

Очевидно, что такой подход (когда задается процент выбросов) можно применить и для обычного цензурирования выборки. Исходя из того факта, что вероятность, с которой случайная величина X (выборка является реализацией этой случайной величины) принимает значения из интервала $(X(i-1), X(i))$ (i -и и $X(i)$ — порядковые статистики), не зависит от распределения и всегда равна $1/(n+1)$ (n — объем выборки), задания процента выбросов, например 5%, можно интерпретировать как построение доверительного интервала, содержащего примерно 95% основной массы значений случайной величины X . Однако на практике такой подход используется относительно редко.

	A	B	C	D	E	F	G
1	Исходная	Отсортированная			Ранг нижнего	Ранг верхнего	
2	выборка	выборка	Ранги	Объем	квартilea	квартilea	
3	1,307219	-1,2	1	55	14,5	41,5	
4	5,278843	-0,5	2		Нижний квартиль	Верхний квартиль	
5		0,168108128	3		0,564890532	2,618222482	
6	0,600888	0,183779397	4		Нижняя граница	Верхняя граница	
7	0,305508	0,262433456	5		-2,515107392	5,698220406	
8	1,520162	0,305507858	6				
48	0,66479	3,707447283	48				
49	0,776932	5,196221693	49				
50	0,428378	5,278842789	50				
51	-0,5	6,026097045	51				
52	-1,2	6,590902832	52				
53	8		53				
54	9		54				
55	10		55				

Рис. 8.9. Результаты цензурирования

	B	C	D	E	F	G
1	Отсортированная	Ранги	Новая		Объем	55 = СЧЕТ (A2:A56)
2	выборка		выборка		Процент винзоризации	5%
3	-1,2	1	-0,5		Нижний ранг	1 = ЦЕЛОЕ (F1*F2/2)
4	-0,5	2	-0,5		Верхний ранг	54 = F1-F3
5	0,168108128	3	0,168108			
6	0,183779397	4	0,183779			
7	0,262433456	5	0,262433	=ЕСЛИ (C2:C56<=F3; ИНДЕКС (B2:B56; F3+1; 1) ;		
8	0,305507858	6	0,305508	ЕСЛИ (C2:C56>=F4; ИНДЕКС (B2:B56; F4; 1) ; B2:B56))		
9	0,340582811	7	0,340583			
10	0,341463856	8	0,341464			
51	5,278842789	50	5,278843			
52	6,026097045	51	6,026097			
53	6,590902832	52	6,590903			
54	8	53	8			
55	9	54	9			
56	10	55	9			

Рис. 8.10. 5% винзоризации

2. Строится винзоризированная выборка $\{x_i^*\}$ по следующей схеме: $x_i^* = x_i$, если $t_n \leq x_i \leq t_n$; $x_i^* = t_n$, если $x_i \leq t_n$; $x_i^* = t_n$, если $x_i \geq t_n$.
3. По винзоризированной выборке $\{x_i^*\}$ вычисляются новые значения среднего \bar{x}^* (по обычной формуле) и выборочная дисперсия s_n^{2*} по формуле

$$s_n^{2*} = \left(\frac{n}{m} \right)^2 \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \text{ где } m - \text{количество неизмененных выборочных значений.}$$
4. Повторяется процесс винзоризации для выборки $\{x_i^*\}$, описанный в п. 2. Если новых винзоризированных выборочных значений не появилось, то на этом процесс завершается. В противном случае повторяется п. 3 и т.д.

	В	С	Д	Е
1	Отсортированная выборка	Новая Ранги	выборка	Объем
2	-1,2	1	0,168108	55 =СЧЕТ (А2:А56)
3	-0,5	2	0,168108	Процент винзоризации 10%
4	0,168108128	3	0,168108	Нижний ранг 2 =ЦЕЛОЕ (F1*F2/2)
5	0,183779397	4	0,183779	Верхний ранг 53 =F1-F3
6	0,262433456	5	0,262433	{=ЕСЛИ (С2:С56<=F3; ИНДЕКС (В2:В56; F3+1; 1); ЕСЛИ (С2:С56>=F4; ИНДЕКС (В2:В56; F4; 1); В2:В56)) }
7	0,305507858	6	0,305508	
8	0,340582811	7	0,340583	
9	0,341463856	8	0,341464	
14	5,278842789	50	5,278843	
52	6,026097045	51	6,026097	
53	6,590902832	52	6,590903	
54	8	53	8	
55	9	54	8	
56	10	55	8	
57				

Рис. 8.11. 10% винзоризации

На практике, если процесс не завершился после двух-трех итераций, его останавливают и увеличивают значение s , после чего винзоризация выполняется сначала. Существуют более сложные модификации описанной процедуры, для которых доказана сходимость за конечное число шагов.

На рис. 8.12 показана реализация первых двух этапов описанной процедуры. Обратите внимание, что количество неизменных значений (ячейка F8) вычисляется как формула массива. Винзоризировано пять наибольших значений выборки.

	В	С	Д	Е
1	Отсортированная выборка	Новая выборка		Среднее
2	-1,2	-1,2		1,956319 =СРЗНАЧ (В2:В56)
3	-0,5	-0,5		Стандартное отклонение 2,341077 =СТАНДОТКЛОНА (В2:В56)
4	0,168108128	0,16811		Константа s 1,5
5	0,183779397	0,18378		Нижняя граница -1,5553 =F\$1-F\$3*F\$2
6	0,262433456	0,26243		Верхняя граница 5,467934 =F\$1+F\$3*F\$2
7	0,305507858	0,30551		
8	0,340582811	0,34058	Число неизменных значений	50
9	0,341463856	0,34146		
10	0,361072452	0,36107		{=СЧЕТ (ЕСЛИ (В2:В56>=F5; ЕСЛИ (В2:В56<=F6; В2:В56; "") ; "")) }
11	0,386983558	0,38698		
12	0,428378072	0,42838		
13	0,442992058	0,44299		{=ЕСЛИ (В2:В56<F5; F5; ЕСЛИ (В2:В56>F6; F6; В2:В56)) }
50	5,196221693	5,19622		
51	5,278842789	5,27884		
52	6,026097045	5,46793		
53	6,590902832	5,46793		
54	8	5,46793		
55	9	5,46793		
56	10	5,46793		

Рис. 8.12. Реализация первых двух этапов винзоризации выборки

Поскольку предполагается, что процедура винзоризации будет иметь не более трех итераций, для ее реализации можно обойтись без циклических вычислений, создав вычисления для трех итераций путем простого копирования формул и их небольшой подгонки под новые диапазоны ячеек, содержащих последовательные винзоризированные выборки. Рабочий лист, рассчитанный на выполнение трех итераций, показан на рис. 8.13. Формулы, по которым проводятся вычисления, аналогичны показанным на рис. 8.12, за исключением стандартного отклонения, которое рассчитывается по формуле, приведенной в п. 3 описания процедуры. Как видно на рис. 8.13, по числу винзоризированных значений процесс сошелся после второй итерации. Если бы этого не произошло, можно было бы изменить значение константы c в ячейке G3. После этого три итерации процесса винзоризации повторились бы автоматически. Винзоризированные значения визуально выделяются с помощью условного форматирования.

	B	C	D	E	F	
1	Отсортированная выборка	Новая выборка 1	Новая выборка 2	Новая выборка 3		
2	-1,2	-1,2			Среднее	1,958319
3	-0,5		-0,5	-0,5	Стандартное отклонение	2,341077
4	0,168108128	0,1681081	0,16810813	0,168108128	Константа c	1,5
5	0,183779397	0,1837794	0,1837794	0,183779397	Нижняя граница	-1,5553
6	0,262433458	0,2624335	0,26243346	0,262433458	Верхняя граница	5,467934
7	0,305507858	0,3055079	0,30550786	0,305507858	Число неизмененных значений	50
8	0,340582811	0,3405828	0,34058281	0,340582811	Итерация 1	
9	0,341463856	0,3414639	0,34146386	0,341463856	Среднее	1,733095
10	0,361072452	0,3610725	0,36107245	0,361072452	Стандартное отклонение	1,928567
11	0,386983558	0,3869836	0,38698356	0,386983558	Нижняя граница	-1,15978
12	0,428378072	0,4283781	0,42837807	0,428378072	Верхняя граница	4,625945
13	0,442992058	0,4429921	0,44299206	0,442992058	Число неизмененных значений	47
14	0,452771365	0,4527714	0,45277137	0,452771365	Итерация 2	
15	0,544177475	0,5441775	0,54417748	0,544177475	Среднее	1,635043
16	0,585603589	0,5856036	0,58560359	0,585603589	Стандартное отклонение	1,812148
17	0,600888197	0,6008882	0,60088819	0,600888197	Нижняя граница	-1,09319
18	0,636547324	0,6365473	0,63654732	0,636547324	Верхняя граница	4,353285
19	0,664789506	0,6647895	0,66478951	0,664789506	Число неизмененных значений	48
20	0,677800958	0,677801	0,67780096	0,677800958	Итерация 3	
21	0,720432989	0,720433	0,72043299	0,720432989	Среднее	1,60173
22	0,768338519	0,7683386	0,76833852	0,768338519	Стандартное отклонение	1,694631
23	0,776932356	0,7769324	0,77693236	0,776932356	Нижняя граница	-0,94022
24	0,781639808	0,7816399	0,78163981	0,781639808	Верхняя граница	4,143877
25					Число неизмененных значений	48

Рис. 8.13. Три итерации винзоризации

Отметим, что процесс винзоризации сошелся по числу винзоризированных значений, но не по значениям среднего и выборочной дисперсии (или, что то же самое, по значениям границ t_n и t_p). Чтобы достигнуть такой сходимости, следовало бы продолжить процесс винзоризации. Однако для описанной процедуры такая сходимость не гарантирована. Поэтому на практике ограничиваются сходимостью по числу винзоризированных значений.

Отметим, что выборки с "обрезанными" экстремальными значениями могут появиться не только в результате винзоризирования, но и "естественным" путем. Например, если выборку составляют наблюдения за некоторой физической переменной, значения которой фиксируются с помощью прибора и этот прибор имеет определенные пределы измерений, то значения физической переменной,

выходящие за эти пределы, будут зафиксированы на уровне предела измерения прибора. Другой пример из эконометрики: обычно при исследовании доходов населения фиксируются точно только доходы, которые лежат в определенных границах. Для доходов, которые меньше определенного уровня (например, ниже уровня бедности) либо больше некоторого другого фиксированного уровня, подсчитывается только их количество, без записи конкретного значения. Такие выборки можно рассматривать как винзоризированные.

8.2. Преобразование данных

Перед выполнением статистического анализа часто проводится преобразование данных. Делается это по нескольким причинам. Во-первых, для того, чтобы распределение преобразованных выборочных значений было свободно от параметров либо было близко к известному распределению (чаще всего — к нормальному), либо имело легко проверяемые свойства. Например, если выборочное распределение явно асимметрично, имеет большой "хвост" вправо и все выборочные значения положительны, то применение логарифмического преобразования приведет к более симметричному распределению, поскольку оно растянет шкалу в окрестности нуля. Во-вторых, необходимость преобразования возникает тогда, когда параметры распределения зависят один от другого (обычно предполагается, что по крайней мере первые моменты — математическое ожидание и дисперсия — не связаны между собой). Например, математическое ожидание и дисперсия пуассоновского распределения совпадают. В подобном случае нужно подобрать такое преобразование, чтобы параметры распределения преобразованных данных были независимы. Отметим, что при таком преобразовании часто улучшаются свойства распределения преобразованных данных, например оно приводит к распределению, близкому к нормальному, или стабилизирует дисперсию выборки (делает ее менее чувствительной к объему выборки и другим параметрам выборки).

Рассмотрим часто используемые преобразования данных.

8.2.1. Преобразование квадратного корня

Это преобразование применяют к распределениям, дисперсия которых совпадает с математическим ожиданием или пропорциональна ему. У преобразованной случайной величины эти параметры можно считать независимыми, при этом ее дисперсия приблизительно равна $1/4$ (или $\hat{\sigma}^2/4$, если дисперсия пропорциональна математическому ожиданию с коэффициентом пропорциональности k). Кроме того, данное преобразование часто приводит к распределению, которое ближе к нормальному, чем исходное. Покажем использование этого преобразования для пуассоновского и χ^2 распределений.

Пуассоновское распределение

Дисперсия случайной величины X , распределенной по закону Пуассона, равна ее математическому ожиданию θ (см. раздел 1.4.4). Простейшим преобразованием этой величины будет \sqrt{X} . При $\theta \leq 4$ более эффективным считается преобразование вида

величины будет $\sqrt{X + \frac{3}{8}}$. При $\theta \leq 4$ более эффективным считается преобразование вида $\sqrt{X + \frac{3}{8}}$. При малых θ иногда рекомендуют использовать преобразование $\sqrt{X} + \sqrt{X+1}$. При малых θ иногда рекомендуют использовать преобразование $\sqrt{X} + \sqrt{X+1}$.

На рис. 8.14 показан рабочий лист, в столбце А которого содержится выборка, имеющая распределение Пуассона с параметром $\theta = 2$ (100 выборочных значений получены с помощью средства Генерация случайных чисел, диапазон выборочных значений назван Выборка). В столбцах В, С и D записаны выборочные значения, преобразованные по формулам \sqrt{X} , $\sqrt{X+3/8}$ и $\sqrt{X} + \sqrt{X+1}$ соответственно. В столбце F вычислены средние и дисперсии (по стандартным формулам с использованием функций СРЗНАЧ и ДИСП), в столбце G — оценки параметра θ (формулы для вычислений показаны на рис. 8.14). Как видно, наилучший результат по значению дисперсии (по близости к значению 0,25) дает преобразование $\sqrt{X+3/8}$, по близости оценки θ к истинному значению — преобразование $\sqrt{X} + \sqrt{X+1}$. На рис. 8.15 показаны гистограммы для первоначальной выборки и для значений, преобразованных по формулам \sqrt{X} и $\sqrt{X+3/8}$ (последняя формула приводит к более симметричному распределению, чем формула \sqrt{X}).

F6		=ДИСП(Формула_1)					
	А	В	С	Д	Е	Г	Тетта
1	Выборка	Формула 1	Формула 2	Формула 3		Выборка	
2	0	0	0,6123724	1	Среднее	2,16	2,16
3	3	1,7320508	1,8371173	3,7320508	Дисперсия	2,4387879	=СРЗНАЧ(Выборка)
4	1	1	1,1726039	2,4142136		Формула 1	
5	0	0	0,6123724	1	Среднее	1,3070342	=F5^2
6	1	1	1,1726039	2,4142136	Дисперсия	0,4562239	
7	1	1	1,1726039	2,4142136		Формула 2	
8	1	1	1,1726039	2,4142136	Среднее	1,5051308	=3/8+F8^2
9	3	1,7320508	1,8371173	3,7320508	Дисперсия	0,2723045	
10	4	2	2,0916501	4,236068		Формула 3	
11	3	1,7320508	1,8371173	3,7320508	Среднее	3,0283621	=(F11^2)-1/4
12	2	1,4142136	1,5411035	3,1462644	Дисперсия	1,2450863	
13	0	0	0,6123724	1			
14	2	1,4142136	1,5411035	3,1462644		[=КОРЕНЬ(Выборка)]	
15	0	0	0,6123724	1			
16	0	0	0,6123724	1		[=КОРЕНЬ(Выборка+3/8)]	
17	0	0	0,6123724	1			
18	4	2	2,0916501	4,236068		[=КОРЕНЬ(Выборка)+КОРЕНЬ(Выборка+1)]	
19	2	1,4142136	1,5411035	3,1462644			
20	3	1,7320508	1,8371173	3,7320508			
21	4	2	2,0916501	4,236068			

Рис. 8.14. Преобразование квадратного корня

Распределение χ^2

Преобразование квадратного корня для выборочных значений, имеющих распределение χ^2 , на практике применяется относительно редко. Это преобразование имеет скорее теоретическое значение и используется для аппроксимации этого распределения нормальным, например, при построении доверительных интервалов. Кроме того, это преобразование дает удовлетворительную аппроксимацию только при достаточно большом значении n степени свободы распределения χ^2 .

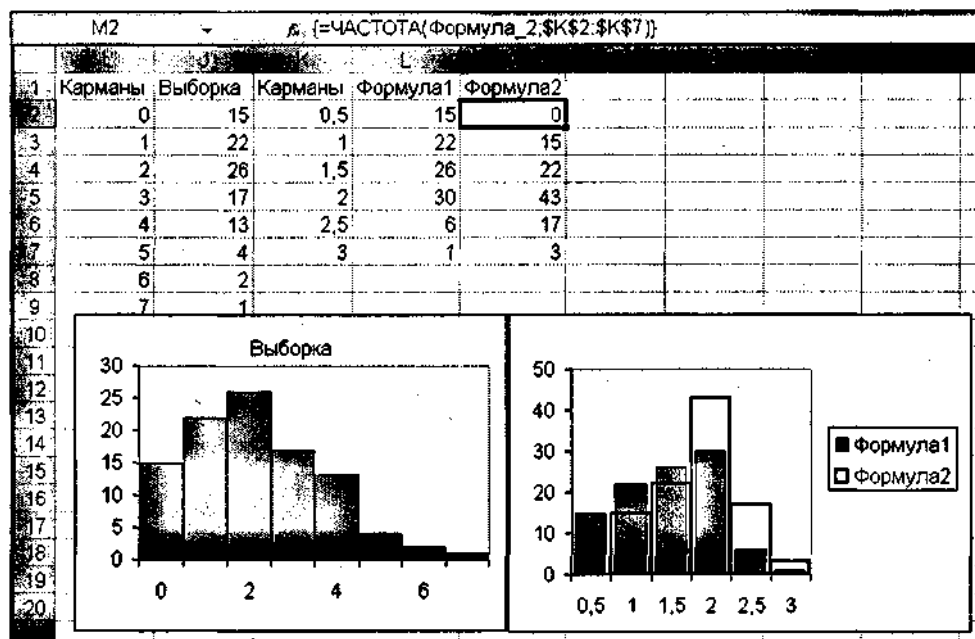


Рис. 8.15. Гистограммы первоначальной и преобразованных выборок

Если $n \geq 30$, то случайная величина $Y = \frac{\sqrt[3]{X/n + 2/9n} - 1}{\sqrt{2/9n}}$ имеет приблизительно нормальное стандартное распределение, здесь X — случайная величина, распределенная по закону χ^2 с n степенями свободы. Если $n \geq 100$, то используется более простое преобразование, которое также приводит к приблизительно нормальному стандартному распределению: $Y = \sqrt{2X} - \sqrt{2n-1}$.

На рис. 8.16 показан рабочий лист, в столбце А которого содержится выборка, имеющая распределение χ^2 со значением степени свободы $n = 50$ (200 выборочных значений получены с помощью формулы массива $\{=\text{ХИ2ОБР}(\text{СЛЧИС}();50)\}$). В столбцах В и С записаны выборочные значения, преобразованные по формулам

$$Y = \sqrt{2X} - \sqrt{2n-1} \text{ и } Y = \frac{\sqrt[3]{X/n + 2/9n} - 1}{\sqrt{2/9n}} \text{ соответственно (обозначены как Формула 1}$$

и Формула 2; диапазонам ячеек, содержащим эти значения, присвоены такие же имена). Формулы преобразования показаны на рис. 8.16. В столбце Е вычислены средние и дисперсии (по стандартным формулам с использованием функций СРЗНАЧ и ДИСП). На рис. 8.17 показаны гистограммы для первоначальной выборки и преобразованных выборок. Как видно, обе формулы в данном случае дают примерно одинаковые результаты: формула 1 дает чуть лучшие значения среднего и выборочной дисперсии, формула 2 — более симметричную гистограмму.

8.2.2. Логарифмическое преобразование

Это преобразование, по-видимому, чаще всего используется на практике, особенно при анализе экономических данных, которые часто имеют логарифмически нормальное или приблизительно логнормальное распределение. Логарифмическое

преобразование также применяют, когда в распределении случайной величины X математическое ожидание и среднеквадратическое отклонение пропорциональны (например, с коэффициентом пропорциональности h). Тогда случайная величина $Y = \ln(X)$ будет иметь дисперсию, приблизительно равную k^2 , т.е. приходим к распределению с почти независимыми математическим ожиданием и дисперсией. Если случайная величина X может принимать нулевые значения, то используется формула $Y = \ln(X + 1)$.

C2		=(((Выборка/50)^(1/3))-1-2/(9*50))/КОРЕНЬ(2/(9*50)))					
	A	B	C	D	E	F	
1	Выборка	Формула 1	Формула 2	Выборка			
2	61,72352	1,16079832	1,15773312	Среднее	49,5596077	=СРЗНАЧ(Выборка)	
3	39,33261	-1,0805533	-1,0864456	Дисперсия	99,7517025	=ДИСП(Выборка)	
4	46,28572	-0,3284695	-0,3143585	Формула 1			
5	56,87249	0,71525471	0,72463233	Среднее	-0,0454381	=СРЗНАЧ(Формула_1)	
6	35,84801	-1,4825244	-1,5080663	Дисперсия	1,02649042	=ДИСП(Формула_1)	
7	46,56822	-0,2991525	-0,284677	Формула 2			
8	61,01653	1,0969827	1,09605987	Среднее	-0,0465273	=СРЗНАЧ(Формула_2)	
9	62,31707	1,21409177	1,20914701	Дисперсия	1,04277719	=ДИСП(Формула_2)	
10	40,41086	-0,9597844	-0,9610532				
11	52,24518	0,27217786	0,28790512				
12	53,77824	0,42106933	0,43535406	=(((Выборка/50)^(1/3))-1-2/(9*50))/КОРЕНЬ(2/(9*50)))			
13	62,22081	1,20546633	1,20083133				
14	46,52241	-0,3039	-0,2894814				
15	58,76694	0,89142986	0,89659494	=КОРЕНЬ(2*Выборка)-КОРЕНЬ(2*50-1))			
16	65,15882	1,46580108	1,45088278				
17	49,89147	0,03926708	0,05580815				

Рис. 8.16. Преобразование исходной выборки

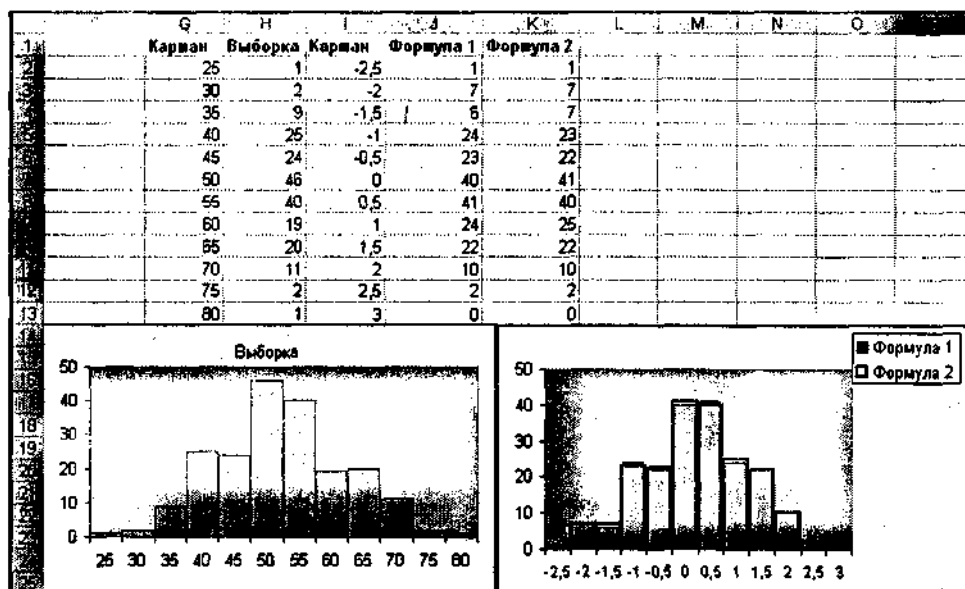


Рис. 8.17. Гистограммы первоначальной и преобразованных выборок

На рис. 8.18 показан рабочий лист, в столбце А которого содержится выборка, имеющая логнормальное распределение с параметрами $m = 1$ и $\sigma^2 = 4$ (1200 выборочных значений получены с помощью формулы массива $\{=\text{ЛОГНОРМОБР}(\text{СЛЧИС}();1;2)\}$). В столбце В записаны выборочные значения, преобразованные по формуле $Y = \ln(X)$. В столбце С вычислены средние и стандартное отклонение преобразованной выборки. На этом же рисунке показана гистограмма для преобразованной выборки.

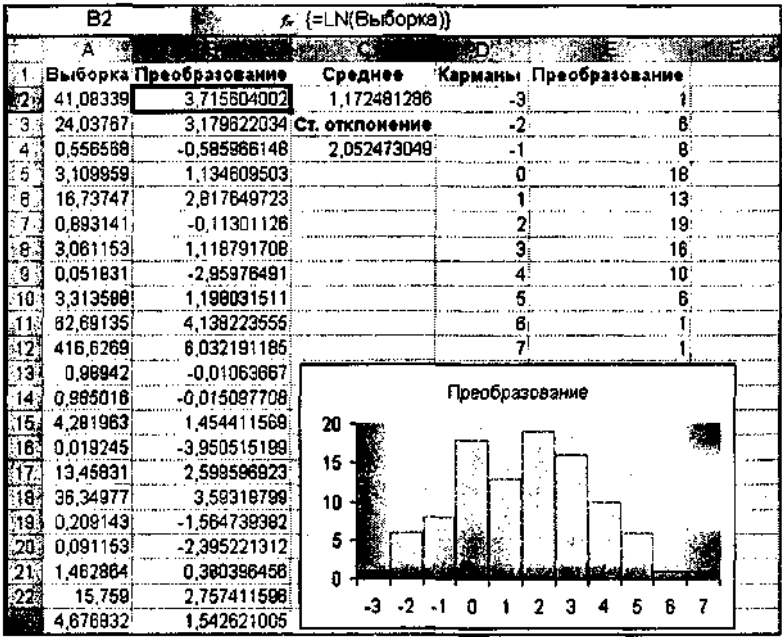


Рис. 8.18. Преобразование исходной выборки

8.2.3. Стандартизирующее преобразование

Если случайная величина X имеет известное математическое ожидание m и дисперсию σ^2 , то случайная величина $Y = X - m$ называется центрированной, величина $Y = X/\sigma$ — нормированной, а $Y = (X - m)/\sigma$ — стандартизированной. Последнее преобразование называется стандартизирующим и используется на практике для получения стандартизированных выборок (в качестве значений m и σ обычно берутся выборочные среднее и стандартное отклонение). Для выполнения этого преобразования в Excel есть специальная функция НОРМАЛИЗАЦИЯ (см. раздел 4.12.2). Отметим, что стандартизирующее преобразование не изменяет тип распределения, а изменяет только значения математического ожидания и дисперсии.

8.3. Построение гистограмм, полигонов и эмпирических функций распределения

На предварительном этапе статистического анализа, как правило, строятся гистограммы, полигоны и эмпирические функции распределения. Это удобный

способ визуального представления статистических данных, который позволяет делать выводы о распределении наблюдаемой случайной величины, реализацией которой является имеющаяся выборка.

8.3.1. Построение гистограммы и эмпирической функции распределения для дискретных случайных величин

Случайная величина, подчиняющаяся дискретному распределению, может принимать конечное или счетное множество значений. Естественно, в конечной выборке всегда есть только конечное количество различных значений. Обычно подобная выборка имеет вид таблицы, в которой указывается, сколько раз каждое значение встречается в выборке. Такая таблица называется *частотной*.

Здесь необходимо сделать небольшое отступление об используемых терминах. Если дискретная случайная величина X принимает значения x_1, x_2, \dots, x_m и данные значения встречаются в выборке соответственно f_1, f_2, \dots, f_m раз, то эти числа называются *частотами* значений X . Значения частот, деленные на объем выборки и выраженные в долях единицы или в процентах, называются *частостями*, *относительными частотами* или *статистическими вероятностями*. *Накопленными частотами* s_i называются количества выборочных значений, не превышающих x_i . Эти же величины, деленные на объем выборки, называются *относительными накопленными частотами* или *накопленными частостями*.

Возвращаемся к выборочным значениям дискретной случайной величины. По частотной таблице построить гистограмму не представляет особых трудностей (за исключением тех случаев, когда значения распределены неравномерно на оси X ; см. ниже). Но иногда, если выборка состоит из последовательных наблюдений, данные не сгруппированы и необходимо подсчитать частоты разных значений. Если выборка небольшого объема и известны значения, которые принимает случайная величина, то сделать это относительно несложно. Однако для больших выборок и особенно в случае, когда неизвестны все значения, принимаемые случайной величиной, задача усложняется. Покажем, как ее можно выполнить в Excel в самом общем случае.

На рис. 8.19 показана выборка из 100 значений, сгенерированная с помощью средства Генерация случайных чисел из пакета анализа с типом распределения Дискретное. Распределение задано в интервале I2:J8 (значения этого распределения получены с помощью функции СЛЧИС).

Анализ выборки следует начать с подсчета количества выборочных значений, для чего применить простую формулу =СЧЁТ(Выборка) (здесь диапазону ячеек, содержащему выборку, присвоено имя Выборка). Далее надо подсчитать количество *различных* значений в выборке. В общем случае это нетривиальная задача. Для ее выполнения можно использовать формулу массива

$$\{=\text{СУММ}(1/\text{СЧЁТЕСЛИ}(\text{Выборка};\text{Выборка}))\}.$$

Данная формула сначала создает виртуальный массив, содержащий для каждого выборочного значения количество таких значений (это делает часть формулы СЧЁТЕСЛИ(Выборка;Выборка)). Например, число 4,56 встречается в выборке 24 раза. Тогда каждый элемент виртуального массива, соответствующий выборочному значению 4,56, будет равен 24. Часть формулы $1/\text{СЧЁТЕСЛИ}(\text{Выборка};\text{Выборка})$ создает новый виртуальный массив, содержащий величины, обратные значениям первого виртуального массива. Например, 24 элемента этого массива, соответствующие выборочному значению 4,56, будут содержать число 0,041667 ($=1/24$). Функция СУММ суммирует значения второго виртуального массива (сумма значений,

соответствующих выборочному значению 4,56, даст 1), и в результате получается искомое количество различных выборочных значений. В рабочем листе, показанном на рис. 8.20, эта формула записана в ячейке С2.

	А	Н	І	Ј
1	Выборка		Распределение	
2	4,56	1,71	0,2	
3	1,71	1,98	0,1	
4	4,44	4,56	0,25	
5	2,48	4,44	0,1	
6	2,48	4,5	0,1	
7	4,1	2,48	0,2	
8	1,71	4,1	0,05	
9	4,56			
10	2,48			
11	1,71			
12	1,98			
13	1,71			
14	1,71			
15	1,71			
16	1,98			
17	1,71			
18	1,98			
19	4,56			
20	4,44			
21	4,56			

Рис. 8.19. Выборка и ее распределение

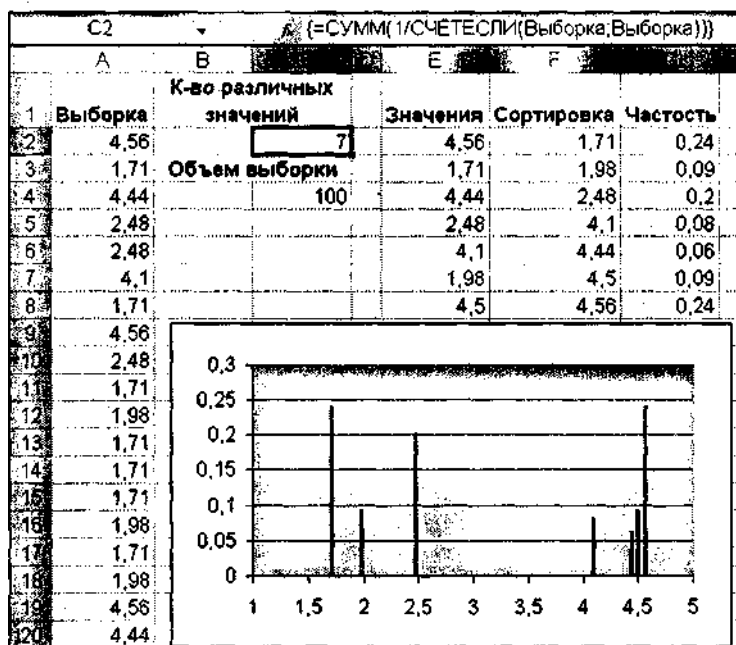


Рис. 8.20. Построение гистограммы для дискретного распределения

Теперь необходимо создать массив, содержащий все различные выборочные значения. Размер такого массива равен числу, подсчитанному предыдущей формулой. На рис. 8.20 этот массив содержится в диапазоне E2:E8 в столбце Значения. Значения этого диапазона вычислены с помощью следующей формулы массива:

{=ИНДЕКС(Выборка;НАИМЕНЬШИЙ(ЕСЛИ(ПОИСКПОЗ(Выборка;Выборка;0)=СТРОКА(ДВССЫЛ("1:"&ЧСТРОК(Выборка)));ПОИСКПОЗ(Выборка;Выборка;0);""));СТРОКА(ДВССЫЛ("1:"&ЧСТРОК(Выборка))))}.

Эта формула, как и предыдущая, взята из книги [20]. Мы не будем описывать, как она работает (это значительно увело бы в сторону от нашей темы); отметим только, что работает она безукоризненно на любых выборках.

Единственный недостаток данной формулы состоит в том, что полученный массив не упорядочен. Однако отметим, что если исходная выборка отсортирована в порядке убывания или возрастания, то данный массив также будет упорядочен. Отсортировать этот массив на месте не удастся, поскольку он получен с помощью формулы массива. Простой выход из такой ситуации заключается в том, чтобы скопировать его в соседний диапазон ячеек и заменить формулы значениями (команда Правка^Специальная вставка, опция Значения). Теперь можно применить стандартную сортировку, для чего следует выделить диапазон и выбрать команду Данные^Сортировка. Если после этого появляется диалоговое окно Обнаружены данные вне указанного диапазона, то в этом окне необходимо установить переключатель Сортировать в пределах указанного выделения и затем щелкнуть на кнопке Сортировка. Упорядоченный по возрастанию массив уникальных выборочных значений на рис. 8.20 показан в столбце F, озаглавленном Сортировка. В столбце G вычислены частоты выборочных значений с применением формулы массива **{=ЧАСТОТА(Выборка;F2:F8)/C4}**. На основании этих данных далее строится гистограмма.

Обычно для построения гистограмм в Excel используется тип диаграммы Гистограмма. Однако этот тип диаграммы располагает данные по оси X равномерно, что вполне подходит, если случайная величина принимает равноотстоящие значения на каком-либо интервале. В нашем случае значения, принимаемые случайной величиной, распределены не равномерно. В такой ситуации можно применить тип диаграммы Точечная и в качестве столбцов гистограммы использовать планки погрешностей, как описано в разделе 6.2.3. Напомним кратко, как построить гистограмму в данном случае.

Сначала строим диаграмму типа Точечная без линий, соединяющих точки данных. Затем выделяем ряд данных и выбираем команду Формат^Выделенный ряд. В открывшемся диалоговом окне Формат ряда данных на вкладке Y-погрешности задаем планку погрешности типа Минус. В качестве величины погрешности задаем Относительное значение 100% (рис. 8.21). На графике появляются вертикальные столбцы от значений данных до оси X. Теперь остается отформатировать планки погрешностей и значения данных. В результате получаем гистограмму выборки, показанную на рис. 8.20.

При построении эмпирической функции распределения для дискретных случайных величин также возникают некоторые сложности, поскольку такая функция имеет ступенчатый вид, но ни средство построения диаграмм Excel, ни средство Гистограмма из пакета анализа подобные графики строить не могут. В разделе 6.2.3 показано, как все-таки в Excel построить такой график.

Далее в пустую ячейку Н2 введем формулу =НЗ-0,000009, а в ячейку 12 число 0. Формулу из ячейки Н2 скопируем в ячейку Н4, а в ячейку 12 введем формулу =13. Выделим ячейки Н4:14 и скопируем их во все свободные ячейки вниз до строки 14. В ячейку Н16 можно ввести число 5, а в ячейку 15 — число 1 (но это не обязательно). Рабочий лист на данном этапе показан на рис. 8.23.

1	Значения	Сортировка	Частость	Сортировка	Накопленные частости
2	4,56	1,71	0,24	1,709991	0
3	1,71	1,98	0,09	1,71	0,24
4	4,44	2,48	0,2	1,979991	0,24
5	2,48	4,1	0,08	1,98	0,33
6	4,1	4,44	0,06	2,479991	0,33
7	1,98	4,5	0,09	2,48	0,53
8	4,5	4,56	0,24	4,099991	0,53
9				4,1	0,61
10				4,439991	0,61
11				4,44	0,67
12				4,499991	0,67
13				4,5	0,76
14				4,559991	0,76
15				4,56	1
				5	1

Рис. 8.23. Все готово для построения графика

Теперь для построения графика эмпирической функции распределения достаточно построить средствами Excel диаграмму типа Точечная с соединительными линиями без маркеров на основе данных диапазона Н2:116. Готовая отформатированная диаграмма показана на рис. 8.24.

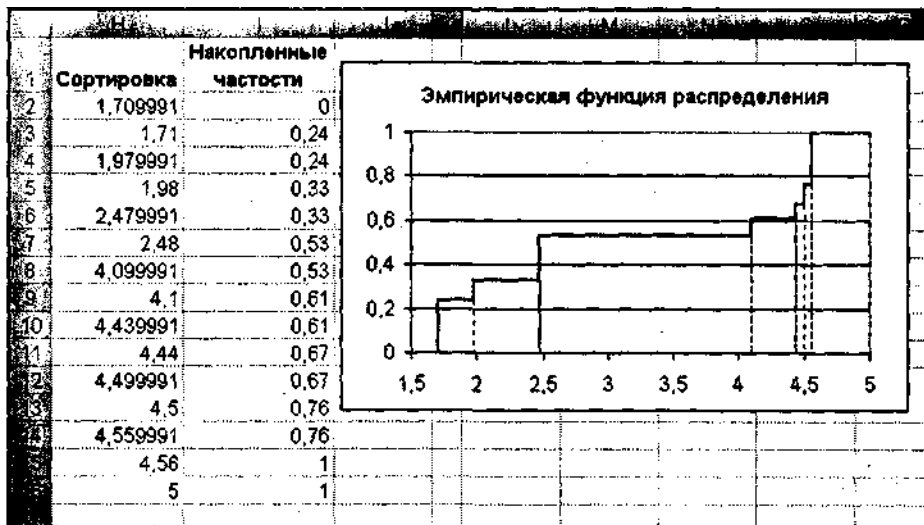


Рис. 8.24. Эмпирическая функция распределения

8.3.2. Построение гистограммы и полигона для непрерывных распределений

Чтобы построить гистограмму для выборки, имеющей непрерывное распределение, необходимо создать для нее частотную таблицу. Для этого сначала вся область изменения выборочных значений разбивается на ряд непересекающихся интервалов и затем подсчитываются количества выборочных значений, попавших в каждый интервал. Такие интервалы часто называют *карманами*, это же название используют функция Excel **ЧАСТОТА** и средство **Гистограмма**. Первая проблема возникает при определении количества таких интервалов, которое, конечно же, должно выбираться в зависимости от объема выборки. В настоящее время наиболее “популярной” формулой, по которой определяется количество k интервалов в зависимости от числа n выборочных значений, является формула Стерджесса: $k = [1 + 3,22 \ln(n)]$ ($[x]$ — целая часть числа x). Для полноты картины приведем другие формулы, рекомендуемые для вычисления k (см., например, [13]).

$k = 10 \lg(n)$, при этом k не должно выходить за интервал $[5, 30]$.

$k = 5 \lg(n)$ и $k \in [6, 20]$.

$k = [3,26 \lg(n) + 0,5] + 1$, если $n \leq 100$; и $k = \min([0,1n], 25) + 1$, если $n > 100$.

$k = [4(0,75(n-1)^2)^{1/5}]$, если $n > 200$; и $k = [0,2n]$, если $n \leq 200$.

$k = \min([\sqrt{n}], 30)$. (В этих формулах $[x]$ — целая часть числа x .)

Какие бы формулы не использовались для вычисления k , следует помнить, что при слишком большом значении k вид распределения искажается случайными значениями частот (поскольку интервалы очень короткие). А при малом числе интервалов сглаживаются и нивелируются характерные особенности распределения (например, наличие двух близкорасположенных мод). Поэтому для качественного анализа строят гистограммы при нескольких значениях k .

После выбора количества интервалов определяется длина интервалов и их границы. Если все интервалы одинаковой длины, то их длина определяется формулой $d = R/k$, где $R = x_{\max} - x_{\min}$ — размах выборки, x_{\max} и x_{\min} — максимальное и минимальное выборочные значения. Часто для того, чтобы минимальное и максимальное значения лежали внутри интервалов, а не на границе, d вычисляют по формуле $d = 1,02R/k$. Если d — дробное число, то за длину интервала принимается или ближайшее целое число, превосходящее d , или ближайшая простая дробь, также не меньшая d . Границы i -го интервала $\Delta_i = [a_{i1}, a_{i2})$ вычисляются по формулам $a_{i1} = a_{11} + (i-1)d$ и $a_{i2} = a_{i1} + id$ ($i = \overline{1, k}$), где a_{11} — нижняя граница интервала Δ_1 . Граница a_{11} равна x_{\min} , если d точно равно R/k либо немного меньше минимального выборочного значения.

Частоты f_i вычисляются как количество выборочных значений, попавших в интервал Δ_i . Обычно в это количество засчитываются значения, которые больше нижней границы интервала или равны ей и меньше верхней границы.

Рассмотрим, как описанные вычисления реализуются в Excel: сначала с помощью формул массивов, а затем с помощью функции **ЧАСТОТА** и средства **Гистограмма**.

Использование формул массивов

На рис. 8.25 показан рабочий лист, в столбце А которого содержатся выборочные значения (этот диапазон ячеек назван Выборка) и вычислены границы интервалов (формулы, по которым выполняются вычисления, также показаны на рис. 8.25). В данном случае выборка имеет равномерное распределение на интервале $[-1, 1]$. Количество интервалов вычисляется в ячейке С8 по формуле Стерджесса. Номера интервалов в столбце D введены как значения арифметической прогрессии с шагом 1 (команда Правка>Заполнить>Прогрессия). Это простейший способ создания интервалов, и он требует выполнения некоторых операций вручную, например, копирования формулы из ячейки Е3 в диапазон Е4:Е9.

C8	=ЦЕЛОЕ(1+3,22*LOG10(C2))						
	A	B	D	E	F	G	H
1	Выборка	Объем выборки	Интервал	Карманы			
2	-0,236	100	1	-1	=ОКРУГЛ((МИН(Выборка)-0,01)/2)		
3	-0,79864	Минимальное значение	2	-0,71	=F2+\$C\$10		
4	0,192969	-0,990050966	3	-0,42	=F3+\$C\$10		
5	0,798212	Максимальное значение	4	-0,13	=F4+\$C\$10		
6	0,769219	0,982482376	5	0,16	=F5+\$C\$10		
7	0,916929	Количество интервалов	6	0,45	=F6+\$C\$10		
8	-0,97101	7	7	0,74	=F7+\$C\$10		
9	-0,18516	Длина интервалов		1,03	=F8+\$C\$10		
10	0,726493	0,29 =ОКРУГЛ(1,02*(МАКС(Выборка)-МИН(Выборка))/D8,2)					
11	-0,72283						
12	-0,50993						

Рис. 8.25. Вычисление границ интервалов

Приведем формулу массива, которая вычисляет границы интервалов, причем нижняя граница первого интервала совпадает с x_{\min} , а верхняя граница последнего — с x_{\max} :

$$\{=\text{МИН}(\text{Выборка})+((\text{СТРОКА}(\text{ДВССЫЛ}("1:"\&(\text{C8}+1))))-1)*(\text{МАКС}(\text{Выборка})-\text{МИН}(\text{Выборка}))/\text{C8}\}$$

Здесь часть формулы $\text{СТРОКА}(\text{ДВССЫЛ}("1:"\&(\text{C8}+1))))-1$ формирует виртуальный массив из целых чисел от 0 до 7. Эти числа затем умножаются на длину интервала, которая вычисляется частью формулы $\text{МАКС}(\text{Выборка})-\text{МИН}(\text{Выборка}))/\text{C8}$. Как видно, для работы данной формулы надо предварительно вычислить только количество интервалов (ячейка С8). Результат использования этой формулы показан на рис. 8.26 в столбце Карманы2.

Небольшое очевидное изменение последней формулы

$$\{=\text{ОКРУГЛ}(\text{МИН}(\text{Выборка})-0,02*\text{Длина};2)+(\text{СТРОКА}(\text{ДВССЫЛ}("1:"\&(\text{C8}+1))))-1)*\text{ОКРУГЛ}(1,04*\text{Длина};2)\}$$

позволяет создавать интервалы, в которых значения x_{\max} и x_{\min} лежат внутри интервалов, а также округляет дробные значения границ интервалов до двух десятичных знаков. Здесь, для того чтобы упростить формулу, длина интервала по формуле $\text{МАКС}(\text{Выборка})-\text{МИН}(\text{Выборка}))/\text{C8}$ вычисляется в ячейке, которой присвоено имя Длина (ячейка С12 на рис. 8.26). Множители 0,02 и 1,04 перед значением Длина надо подбирать таким образом, чтобы вычисленное значение

ОКРУГЛ(МИН(Выборка)-0,02*Длина;2) было меньше x_{\min} . Например, в данном примере при множителе 0,01 значение нижней границы первого интервала было больше x_{\min} . Это результат округления — если не использовать функцию ОКРУГЛ, то любой положительный множитель будет давать значение нижней границы первого интервала, меньшее x_{\min} . Результаты вычислений по последней формуле показаны на рис. 8.26 в столбце Карманы3.

F2		А: {=МИН(Выборка)+((СТРОКА(ДВССЫЛ("1:"&(C8+1)))-1)*(МАКС(Выборка)-МИН(Выборка))/C8)}					
1	Выборка	Объем выборки	Интервал	Карманы	Карманы2	Карманы3	1
2	-0,236	100	1	-1	-0,990051	-1	
3	-0,79864	Минимальное значение	2	-0,71	-0,70826	-0,71	
4	0,192969	-0,990050966	3	-0,42	-0,42647	-0,42	
5	0,798212	Максимальное значение	4	-0,13	-0,14468	-0,13	
6	0,769219	0,982482376	5	0,16	0,1371109	0,16	
7	0,916929	Количество интервалов	6	0,45	0,4189014	0,45	
8	-0,97101	7	7	0,74	0,7006919	0,74	
9	-0,18516	Длина интервалов	1,03			1,03	
10	0,726493	0,29					
11	-0,72283						
12	-0,50993	Длина	0,281790477				
13	-0,90905						
14	-0,93524						

Рис. 8.26. Формулы массивов для вычисления границ интервалов

Теперь подсчитаем количество выборочных значений, попадающих в соответствующие интервалы, т.е. создадим частотную таблицу. Для этого можно использовать формулу массива (границы интервалов записаны в столбце G начиная со второй строки)

$$\{=СУММ((Выборка \geq G2)*(Выборка < G3))\},$$

которая записывается в первую ячейку частотной таблицы, а затем копируется вниз. Здесь в значения частот засчитываются выборочные значения, которые больше нижней границы интервала или равны ей и меньше верхней границы. На рис. 8.27 по этим формулам в столбце H (озаглавленном Частота) подсчитаны частоты для интервалов Карманы3. В столбце I (озаглавленном Частота2) по аналогичным формулам подсчитаны частоты для интервалов Карманы2, где в качестве нижней границы первого интервала взято x_{\min} , а верхней границей последнего — x_{\max} . Как видно на рис. 8.27 в строке состояния, в этом случае сумма частот не равна 100 (т.е. объему выборки), поскольку в последнем интервале не засчитано значение x_{\max} . Таким образом, следует создавать такие интервалы, чтобы значения x_{\min} и x_{\max} находились внутри интервалов.

Для вычисления частотей значения частот необходимо разделить на количество выборочных значений. Если вычислять частоты без предварительного вычисления частот, то для этого можно использовать формулу

$$\{=СУММ((Выборка \geq G2)*(Выборка < G3))/ЧЁТ(Выборка)\},$$

которая записывается в первую ячейку таблицы частотей, а затем копируется вниз (здесь предполагается, что границы интервалов записаны в столбце G начиная со второй строки). На рис. 8.28 значения частотей вычислены по приведенным выше формулам и записаны в столбце I, озаглавленном Частости.

и играет свою роль только в тех случаях, когда границы некоторых интервалов совпадают с выборочными значениями (например, если нижняя граница первого интервала и верхняя граница последнего интервала равны соответственно x_{\min} и x_{\max}).

E2		={ЧАСТОТА(Выборка;D2:D9)}			
	A	B	C	D	E
1	Выборка	Объем выборки	Карманы	1	Частота
2	-0,236		100	-1	0
3	-0,79864	Минимальное значение		-0,71	17
4	0,192969		-0,990050966		16
5	0,798212	Максимальное значение		-0,13	18
6	0,769219		0,982482376	0,16	11
7	0,916929	К-во интервалов		0,45	10
8	-0,97101		7	0,74	15
9	-0,18516	Длина интервалов		1,03	15
10	0,726493		0,29		
11	-0,72283				
12	-0,50993				

Рис. 8.30. Применение функции ЧАСТОТА

Средство пакета анализа Гистограмма описано в разделе 5.2. Отметим, что если не заданы интервалы карманов, то они вычисляются автоматически следующим способом. Вычисляются количество интервалов по формуле Стерджесса $k = [1 + 3,22 \ln(n)]$ (n — объем выборки) и длина интервалов как $d = R/(k - 1)$ (R — размах выборки). Затем последовательно вычисляются границы интервалов, причем за нижнюю границу первого интервала берется x_{\min} . Обращаем внимание, что средство Гистограмма строит на один интервал больше, чем вычисляет формула Стерджесса. Последний интервал Гистограмма обозначает как Еще.

Значения частот средство Гистограмма вычисляет так же, как и функция ЧАСТОТА. На рис. 8.31 показаны интервалы карманов, вычисленные средством Гистограмма, и построенная им гистограмма частот. Отметим, что в качестве подписей к оси X Гистограмма берет значения из массива Карман. Поэтому для того, чтобы изменить формат подписей на диаграмме, следует изменить формат числовых значений в массиве Карман. В остальном эту диаграмму можно форматировать так же, как любую другую диаграмму Excel.

8.4. Вычисление точечных оценок параметров распределения

Вычисление различных оценок параметров распределения предшествует любому более-менее глубокому статистическому анализу имеющихся выборочных данных. Уже на этапе предварительного анализа используются эти оценки, особенно первых моментов, например при цензурировании и преобразовании (нормализации) исходных данных (см. разделы 8.1 и 8.2). Но, прежде всего, оценки параметров дают первоначальное представление о типе и характере распределения выборки (конечно, наряду с другими средствами предварительного анализа, например с гистограммами и эмпирической функцией распределения). Параметры распределения можно разбить на несколько групп.

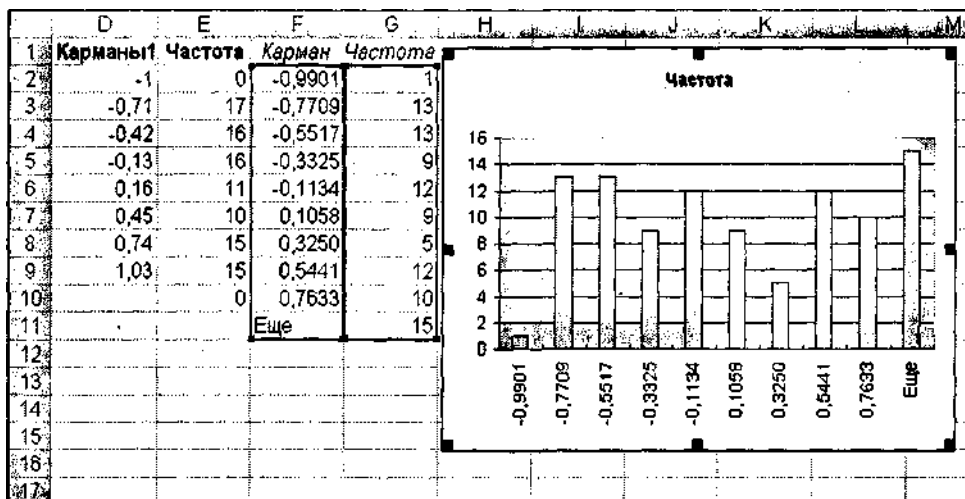


Рис. 8.31. Применение средства Гистограмма

1. **Параметры положения.** Характеризуют положение выборочных данных (точнее, генеральной совокупности) на числовой оси. К таким параметрам можно отнести минимальное и максимальное выборочные значения и выборочные квантили. "Средние значения" местоположения генеральной совокупности характеризуют выборочные средние (арифметическое, геометрическое или гармоническое), медиана и мода.
2. **Параметры разброса.** Характеризуют степень разброса выборочных данных относительно некоторого "среднего значения". К ним, в первую очередь, относятся выборочные дисперсия и среднее квадратическое отклонение, размах выборки и интерквартильный размах (разность между выборочными верхним и нижним квартилями), коэффициент вариации (отношение выборочного среднее квадратического отклонения к среднему) и др.
3. **Параметры формы распределения.** Определяют "геометрические" характеристики распределения, например симметричность и "острота" формы плотности вероятности. К таким параметрам, прежде всего, относятся выборочные коэффициенты асимметрии и эксцесса, а также количество мод (если по гистограмме можно четко определить наличие нескольких мод), относительное расстояние между медианой и средним и т.п.

Для вычисления большинства перечисленных параметров в Excel предусмотрены соответствующие функции (см. главу 4), а также средство Описательная статистика из пакета анализа (см. раздел 5.1). Например, для выборки, которая использовалась в примерах предыдущего раздела, средство Описательная статистика рассчитало статистические показатели, показанные на рис. 8.32.

Приведем список основных точечных оценок параметров распределений с соответствующими формулами и названиями функций Excel, которые вычисляют эти оценки. (В формулах x_i — выборочные значения, n — объем выборки, $x_{(i)}$ — члены вариационного ряда, построенного по исходной выборке.) Также укажем, выполняются ли эти вычисления средством Описательная статистика.

Оценка	Формула	Функция Excel	Описательная статистика
Параметры положения			
Среднее	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$	СРЗНАЧ, СРЗНАЧА	Да
Усеченное среднее	$\bar{x}_a = \frac{1}{n - [an]} \sum_{i=1+[an]/2}^{n-[an]/2} x_i$	УРЕЗСРЕДНЕЕ	Нет
Среднее геометрическое	$\bar{x}_{\text{геом}} = \sqrt[n]{x_1 x_2 \cdot \dots \cdot x_n}$	СРГЕОМ	Нет
Среднее гармоническое	$\bar{x}_{\text{гарм}} = n / \sum_{i=1}^n \frac{1}{x_i}$	СРГАМ	Нет
Минимальное выборочное значение	$x_{\min} = \min(x_1, x_2, \dots, x_n)$	МИН, МИНА, НАИМЕНЬШИЙ	Да
k -е наименьшее значение		НАИМЕНЬШИЙ	Да
Максимальное выборочное значение	$x_{\max} = \max(x_1, x_2, \dots, x_n)$	МАКС, МАКСА, НАИБОЛЬШИЙ	Да
k -е наибольшее значение		НАИБОЛЬШИЙ	Да
Медиана	$m = x_{(k+1)}, \text{ если } n = 2k + 1;$ $m = (x_{(k)} + x_{(k+1)})/2, \text{ если } n = 2k$	МЕДИАНА, КВАРТИЛЬ ¹ , ПЕРСЕНТИЛЬ ²	Да
Мода	(См. раздел 8.4.2)	МОДА	Да
Квантили	(См. разделы 1.2.3 и 4.2.1)	КВАРТИЛЬ	Нет

¹ Эта функция вычисляет медиану при значении аргумента Часть, равном 2.

² Эта функция вычисляет медиану при значении аргумента $k = 0,5$.

Продолжение табл.

Описательная
статистика

Оценка

Формула

Функция Excel

Параметры положения			
Процентили	(См. раздел 1.2.3)	ПЕРСЕНТИЛЬ	Нет
A-й начальный момент	$\bar{m}_k = \frac{1}{n} \sum_{i=1}^n x_i^k$	Специальной функции нет, но легко вычисляется с помощью функции СУММ или СРЗНАЧ	Нет
Параметры разброса			
Выборочная дисперсия	$S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$	ДИСПР, ДИСПРА	Нет
Несмещенная выборочная дисперсия	$s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$	ДИСП,ДИСПА	Да
Усеченная выборочная дисперсия	$\bar{s}_a^2 = \frac{1}{n-[an]-1} \sum_{i=1+[an]/2}^{n-[an]/2} (x_i - \bar{x}_a)^2$	После предварительного цензурирования выборки можно применить функции ДИСП и ДИСПА	Нет
Среднеквадратическое отклонение	$S_n = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$	СТАНДОТКЛОНП, СТАНДОТКЛОНПА	Нет
Стандартное отклонение	$s_n = \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}$	СТАНДОТКЛОН, СТАНДОТКЛОНА	Да
Среднее абсолютное отклонение	$d_n = \frac{1}{n} \sum_{i=1}^n x_i - \bar{x} $	СРОТКЛ	Нет

Оценка

Формула

Функция Excel

Параметры разброса

Коэффициент вариации

$$v = \frac{s}{\bar{x}} \cdot 100\%$$

Очевидная формула с использованием функций СТАНДОТКЛОН и СРЗНАЧ

Нет

Размах

$$R = x_{\max} - x_{\min}$$

Очевидная формула с использованием функций МАКС и МИН

Да
(называется
Интервал)

Интерквартильный размах

$$R_{0,5} = \xi_{0,75} - \xi_{0,25}$$

Очевидная формула с использованием функции КВАРТИЛЬ

Нет

&-й центральный момент

$$\bar{\mu}_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$$

Вычисляется с помощью функций СУММ и СРЗНАЧ

Нет

Параметры формы распределения

Коэффициент асимметрии

$$\hat{\beta}_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_n} \right)^3$$

СКОС

Да

Коэффициент эксцесса

$$\hat{\beta}_2 = \frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^n \left(\frac{x_i - \bar{x}}{s_n} \right)^4 - \frac{3(n-1)^2}{(n-3)(n-3)}$$

ЭКСЦЕСС

Да

	A	B	C	D	E
1	Выборка		Выборка		
2	-0,236		Среднее	-0,029086	
3	-0,79864		Стандартная ошибка	0,0618471	
4	0,192969		Медиана	-0,06766	
5	0,798212		Мода	#ИД	
6	0,769219		Стандартное отклонение	0,6184714	
7	0,916929		Дисперсия выборки	0,3825069	
8	-0,97101		Экссесс	-1,330919	
9	-0,18516		Асимметричность	0,0997568	
10	0,726493		Интервал	1,9725333	
11	-0,72283		Минимум	-0,990051	
12	-0,50993		Максимум	0,9824824	
13	-0,90905		Сумма	-2,908597	
14	-0,93524		Счет	100	
15	-0,67174		Наибольший(1)	0,9824824	
16	-0,56078		Наименьший(1)	-0,990051	
17	-0,96582				

Рис. 8.32. Статистические характеристики выборки, полученные с помощью средства *Описательная статистика*

Приведенные формулы применимы для любых распределений. Для некоторых конкретных распределений существуют специальные точечные оценки параметров распределения, которые будут показаны в главе 10. Отдельного рассмотрения требуют формулы для выборки из дискретной генеральной совокупности, представленной в виде частотной таблицы, а также некоторые пояснения необходимы для оценки моды.

8.4.1. Точечные оценки дискретного распределения

Пусть выборка из дискретной генеральной совокупности представлена в виде частотной таблицы, где для каждого значения x_1, x_2, \dots, x_m указываются соответствующие частоты f_1, f_2, \dots, f_m . Обозначим как n сумму всех частот, т.е.

$n = \sum_{i=1}^m f_i$. Приведем математические формулы и формулы Excel для вычисления

оценок моментов распределения. В формулах Excel будем предполагать, что значения x_1, x_2, \dots, x_m располагаются в диапазоне ячеек с именем *Значения*, значения частот f_1, f_2, \dots, f_m — в диапазоне с именем *Частота*, а значение n — в ячейке с именем *N* (для нахождения n можно применить формулу `=СУММ(Частота)`).

Оценка

Формула

Формула Excel

Выборочное среднее

$$\bar{x} = \frac{1}{n} \sum_{j=1}^m x_j f_j$$

`=СУММПРОИЗВ(Значения;Частота)/N`

A-й начальный момент

$$\mu_k' = \frac{1}{n} \sum_{j=1}^m x_j^k f_j$$

`=СУММПРОИЗВ(Значения^k;Частота)/N`
(значение k записано в ячейке с именем *K*)

Оценка	Формула	Формула Excel
Выборочная дисперсия	1	=СУММПРОИЗВ((Значения-Среднее)^2;Частота)/Ы (значение х записано в ячейке с именем Среднее) ¹
Несмещенная выборочная дисперсия	$s^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$	=СУММПРОИЗВ((Значения-Среднее)^2;Частота)/(М-1) (значение х записано в ячейке с именем Среднее)
Среднеквадратическое отклонение	$s = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$	=КОРЕНЬ(СУММПРОИЗВ((Значения-Среднее)^2;Частота)/М) (значение х записано в ячейке с именем Среднее)
Стандартное отклонение	$\sigma = \sqrt{\frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^2}$	=КОРЕНЬ(СУММПРОИЗВ((Значения-Среднее)^2;Частота)/(М-1)) (значение х записано в ячейке с именем Среднее)
Среднее абсолютное отклонение	$\bar{x}_a = \frac{1}{n} \sum_{i=1}^n x_i - \bar{x} $	=СУММПРОИЗВ(АВ8(Значения-Среднее);Частота)/М (значение х записано в ячейке с именем Среднее)
k-й центральный момент	$\bar{x}_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$	=СУММПРОИЗВ((Значения-Среднее)^К;Частота)/1\1 (значение k записано в ячейке с именем К)
Мода	Значение, которому соответствует наибольшее значение частоты	=ИНДЕКС(Значения;ПОИСКПОЗ(МАКС(Частота);Частота))^2

Для вычисления медианы также можно создать формулу Excel, однако она будет очень громоздкой и неудобной в использовании. Вычисления значительно сократятся и упростятся, если предварительно отсортировать частотную таблицу по возрастанию значений и подсчитать накопленные частоты, а также найти значение x_m , которому соответствует накопленная частота F_m , меньшая $ga/2$, и следующее по величине значение x_{m+1} , которому соответствует накопленная частота P_{m+1} большая или равная $ga/2$. Тогда медиана M вычисляется по формуле

$$M = x_m + \frac{(x_{m+1} - x_m) \cdot (ga/2 - F_m)}{P_{m+1} - F_m}$$

На рис. 8.33 показано вычисление медианы по этой формуле. Значения x_m и x_{m+1} и соответствующие им значения частот и накопленных частот выделены серым цветом.

¹ Для этой формулы сначала необходимо вычислить среднее. Чтобы найти только оценку дисперсии, без промежуточных вычислений, в последнюю формулу вместо Среднее следует вставить вышеприведенную формулу вычисления среднего. Это же замечание относится и к приведенным ниже формулам.

² Если есть группа из нескольких значений, которым соответствуют одинаковые наибольшие значения частот, то эта формула возвращает первое встреченное значение из данной группы.

ЕЗ	* =A5+(A6-A5)*(G1-C5)/C6					
А	Накопленные					
"1	Частоты	частоты	N=	5731	N/2=	287
2	Значения	49!	49!			
3	0,0363	10!	591 Медиана	0.1633		
4	0,1180	98!	157!			
5	0,1550	98	<•••< 2 5 5			
6	0,2391	63	318			
7	0,2688	48	366			
8	0,3338	50	416			
9	0,3833	82	498			
10	0,5621	38	536			
11	0,6043	37	573			

Рис. 8.33. Вычисление медианы

8.4.2. Вычисление моды для непрерывных распределений

Как указывалось при описании функции МОДА (см. раздел 4.11.3), эта функция на самом деле не вычисляет моду распределения (впрочем, как и средство Описательная статистика). Она просто определяет выборочное значение, которое встречается в выборке наиболее часто. Но поскольку для непрерывных случайных величин вероятность принятия одинаковых значений равна нулю, то в выборках, имеющих непрерывное распределение, одинаковые значения практически не встречаются (а если и встречаются, то это, скорее всего, артефакт). На практике мода непрерывных распределений определяется следующим образом.

1. По выборочным значениям строится гистограмма (или полигон) (см. раздел 8.3.2), по виду которой определяется интервал, в котором может находиться мода (такой интервал называется *модальным*). Пусть границами этого интервала служат числа x_m и x_{m+1} .
2. Значение моды определяется по следующей формуле:

$$m = x_m + (x_{m+1} - x_m) \cdot \frac{f_m - f_{m-1}}{f_m - f_{m-1} + f_{m+1} - f_m}$$

где f_m , f_{m-1} и f_{m+1} — частоты соответственно модального, предшествующего модальному и следующего за модальным интервалов. В этой формуле вместо частот можно использовать частоты.

Если определен модальный интервал, то реализация такой формулы в Excel не вызывает затруднений.

Подбор распределения

Определение вида распределения случайной величины X , реализацией которой являются имеющиеся выборочные значения, — одна из основных целей любого статистического анализа. По большому счету, если известно распределение выборки, на этом можно заканчивать статистический анализ одномерной выборки, поскольку известная функция распределения может дать исчерпывающую информацию о случайной величине X . На практике, конечно, распределение выборки неизвестно, — в лучшем случае исходя из каких-либо априорных соображений можно предположить, что это распределение принадлежит какому-нибудь известному *классу распределений*. Но, поскольку любое конкретное распределение определяется некоторым набором параметров, возникает задача, во-первых, проверить гипотезу о том, что распределение данной выборки действительно принадлежит данному классу распределений, а во-вторых, найти числовые значения параметров распределения.

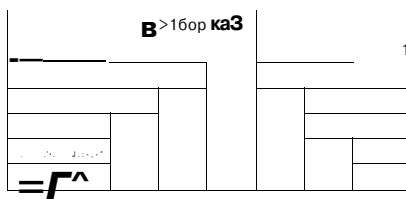
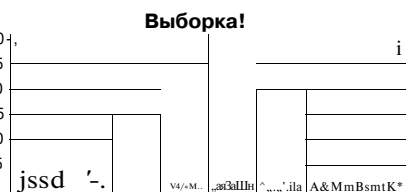
Вторая задача, нахождение числовых параметров распределения, разрешима, если в качестве числовых значений этих параметров принять их статистические оценки, рассчитанные по выборочным значениям. Как правило, данные параметры можно выразить через первые моменты распределения (если параметры, определяющие распределение, сами не являются этими моментами), поэтому и с вычислительной точки зрения оценивание таких параметров является относительно простой задачей. (Конечно, это утверждение справедливо только в том случае, если не включать в задачу оценки параметров проблему надежности и точности полученных оценок.)

В любом случае сначала надо определить класс распределений, к которому может относиться распределение имеющейся выборки. Если не привлекать каких-либо априорных предположений о классе распределений, то остается определить этот класс только на основании выборочных значений, например по виду гистограммы или полигона, либо на основании некоторых выборочных статистик (чаще всего для этого используются выборочные коэффициенты асимметрии и эксцесса). Предварительному определению класса распределений посвящен следующий раздел этой главы. Но далее необходимо проверить выдвинутую гипотезу о том, что выборка действительно имеет распределение из данного класса распределений. Проверка такой гипотезы рассмотрена в последующих двух разделах главы.

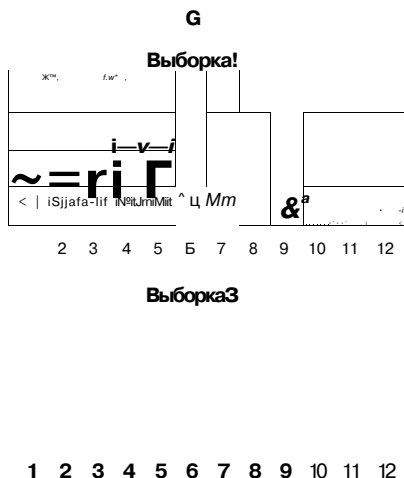
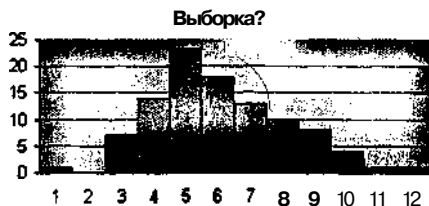
9.1. Предварительное определение класса распределения

Основным "орудием" для предварительного определения класса распределения генеральной совокупности служит гистограмма или полигон частот. Однако

<u>A</u>	<u>j3</u>	<u>'d4c'Sk</u>	
Выборка1	Выборка2	Выборка3	Карманы
-0,3002322	-1,2776832	0,24425731	-- 2,8
1.2764735	1.1983502	1,7331331	•276875
-2.1835876	-0,2341812;	1,09502253	-1,375
-1.0867006	"-0°902042;	-1.6904323	0,6625
-1,8469109	-0,9776295!	-0,7735071	0,05
-2,1179312	-0,5679249;	-0,4040476	0,7625
0,1348531b":.365493!	-0,3269906	1,475
-0.3702405	1.3264616j	-0.0852845	2.1875



Выборка1	Выборка2	Выборка3	Карманы
-0,3002322	-1,2776832	0,24425731	-2,8
1,2764735	" 1.1983502,	1.7331331"	-2.3
-2,1835876	-0,23418121	1.09502253	-1,8
-1,0867006	-0,6902042	-1,6904323	-1,3
-1,8469109	^-0.9776295^	-6.7735071"	" -0.8*
~2.1179312	[-0,5679249^	-0.4040476"	-0,3
0,1348531	-0,365493	-0,3269906^	0,2*
л 3704/1П5			



Глава 9. Подбор распределения 287

Обращаем внимание на несимметричность гистограмм — для любой выборки, имеющей симметричное распределение, гистограмма будет иметь определенный скос в ту или иную сторону. Этот факт теоретически обоснован, но для практического анализа от этого не легче. Если выборка достаточно большого размера, можно попробовать разбить ее на две и для каждой половины построить свою гистограмму. Если гистограммы будут иметь скос в разные стороны, то это может служить “намеком” на симметричность распределения. На основе гистограмм, показанных на рис. 9.1, единственное, что можно утверждать с большой долей уверенности, — что распределение выборки одномодально.

Итак, нужен “опытный глаз”, чтобы на основании гистограмм (или полигонов) сделать выводы о принадлежности распределения выборки тому или иному классу распределений. Чтобы сделать аналогичные выводы на основе *пробит-графиков*, которые мы сейчас рассмотрим, также необходим опыт статистических исследований, но здесь уже возможны и некоторые числовые оценки близости выборочного распределения к некоторому классу распределений.

9.1.1. Построение пробит-графиков

Пробит-график — это график зависимости $y = \Phi^{-1}(F_n(x))$, где F_n — эмпирическая функция распределения, Φ^{-1} — функция, обратная к некой функции распределения. Если распределение выборки совпадает с распределением Φ , то пробит-графиком для такой выборки будет прямая линия. По степени отклонения пробит-графика от прямой линии судят о близости распределения выборки к распределению Φ . Таким образом, для построения пробит-графика необходимо иметь предположение о том, какому классу распределений может принадлежать распределение выборки. Простота построения пробит-графиков, а также числовые показатели отклонения пробит-графика от прямой линии, позволяют просмотреть несколько вариантов предполагаемых функций распределений и выбрать из них наиболее подходящий.

Существует несколько способов построения пробит-графиков¹. Первый способ предполагает выборку большого объема и предназначен именно для подбора типа распределения. Второй способ применяется к малым выборкам и часто используется для определения выбросов (см. раздел 8.1). Рассмотрим первый способ.

На рис. 9.3 показаны выборка (имеющая стандартное нормальное распределение и построенная с помощью средства Генерация случайных чисел), а также таблица частостей и накопленных частостей, рассчитанная по этой выборке. (О создании такой таблицы речь идет в разделе 8.3.2.) Накопленные частости — это эмпирическая функция распределения, график которой также показан на рис. 9.3.

В качестве аргументов x для построения пробит-графика возьмем середины интервалов-карманов. На рис. 9.4 эти значения записаны в столбце Значения x . Теперь осталось подсчитать значения y , вычисляемые по формуле $y = \Phi^{-1}(F_n(x))$. Значения $F_n(x)$ — это значения накопленных частостей, записанные в столбце Накопленные частости. Построим пробит-графики для нормального распределения и равномерного, сосредоточенного на интервале $[-3, 3]$. В первом случае используем функцию НОРМСТОБР (см. раздел 4.7.6). Во втором случае, как не-

¹ Ранее, в докомпьютерную эпоху, для построения пробит-графиков существовала особая вероятностная бумага со специальной шкалой, рассчитанной для разных распределений, в частности для нормального и логнормального.

трудно проверить, функция, обратная к функции распределения, имеет вид $\Phi^{-1}(*) = b(y - 0,5)$. Подсчитанные значения y для первого и второго случаев показаны на рис. 9.4 в столбцах Нормальное y и Равномерное y . Отметим, что крайние значения эмпирической функции распределения (т.е. значения 0 и 1) для вычислений не используются. Причины этого очевидны — если распределение Φ определено на бесконечном интервале, то функция $\Phi^{-1}(l)$ также должна принимать бесконечные значения при $x = 0$ и $x = 1$.

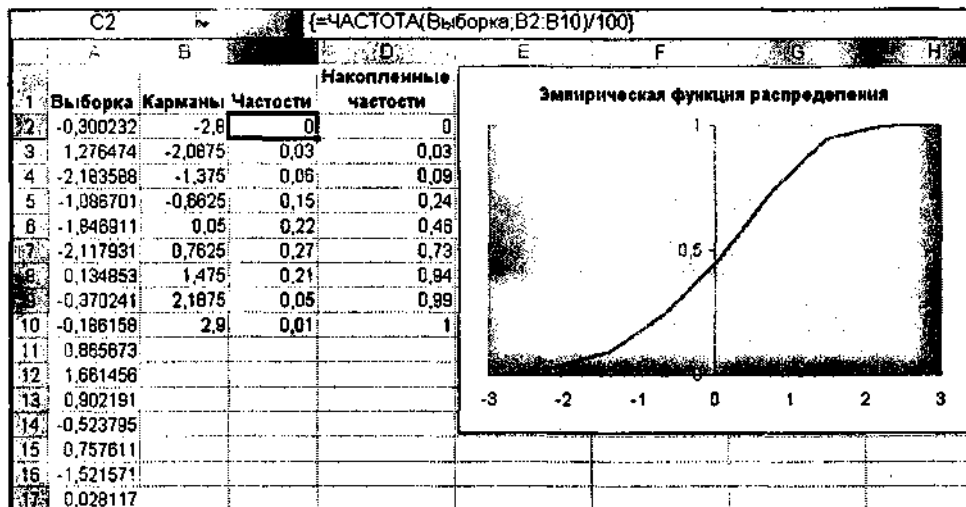


Рис. 9.3. Выборка и ее эмпирическая функция распределения

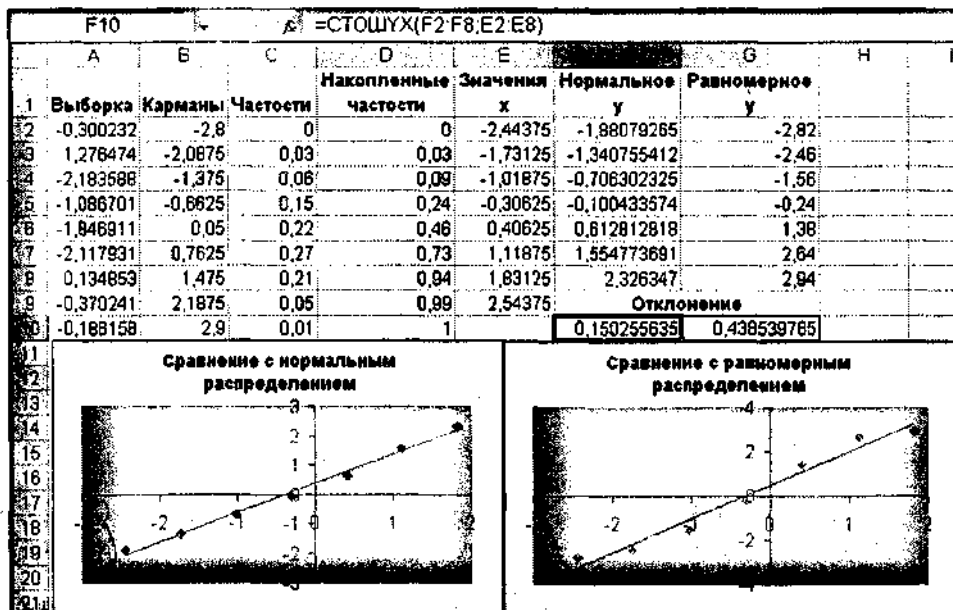


Рис. 9.4. Построение пробит-графиков

Для построения пробит-графиков сначала применяется тип диаграммы Точечная, а затем проводится прямая линейного тренда (см. раздел 6.2.1). Пробит-графики для нашего примера показаны на рис. 9.4. Даже "на глаз" видно, что в данном случае эмпирическая функция распределения ближе к нормальному распределению, чем к равномерному. Но чтобы подтвердить это, можно подсчитать стандартную ошибку приближения, т.е. квадратный корень из средней суммы остатков (см. раздел 3.4.3). Для этого можно использовать функцию Excel

	А	В	С
	Выборка	Вариационный ряд	Функция распределения
1	-0,30023	-2,19358764	0,05
2	1,276474	-2,117931217	0,1
3	-2,19358	-1,646910891	0,15
4	-1,0867	-1,742462709	0,2
5	-1,84691	-1,521570394	0,25
6	-2,11793	-1,086700649	0,3
7	0,134853	-0,523785052	0,35
8	-0,37024	-0,370240514	0,4
9	-0,16618	-0,300232159	0,45
10	0,985673	-0,166157649	0,5
11	1,661456	0,028117029	0,55
12	0,902191	0,134853053	0,6
13	-0,5238	0,595741767	0,65
14	0,757611	0,75761136	0,7
15	-1,52157	0,757713678	0,75
16	0,028117	0,895872973	0,8
17	-1,74248	0,902191459	0,85
18	1,44767	1,27647354	0,9
19	0,757714	1,447670002	0,95
20	0,585742	1,661455826	1

Рис. 9.5. Вычисление значений эмпирической функции распределения

ния $F_n(x)$ при $x = x_{(i)}$ равно $i/(n + 1)$. По этой формуле подсчитаны значения эмпирической функции распределения на рис. 9.5 в столбце Функция распределения. Значения u вычисляются таким же образом, как при первом способе построения пробит-графика для проверяемых распределений. Затем строятся пробит-графики, прямые линейного тренда, а также рассчитываются стандартные ошибки приближения. Пробит-графики для данного примера показаны на рис. 9.6. Здесь визуально сложно определить, к какому распределению ближе эмпирическое распределение. Однако значения стандартных ошибок приближения по-прежнему показывают, что эмпирическую функцию распределения лучше приближает нормальное распределение.

Построенные таким способом пробит-графики часто используются для определения выбросов, — выборочные значения, которые порождают точки, далеко отстоящие от линии тренда, подозрительны как артефакты. Существенно, что здесь можно определить не только экстремальные выбросы, но и выбросы, которые лежат внутри интервала изменения выборочных значений.

STOLUX (см. раздел 4.9.3). Значения стандартной ошибки приближения к нормальному и равномерному распределениям на рис. 9.4 показаны в ячейках F10 и G10 соответственно. Эти значения также показывают, что эмпирическая функция распределения ближе к нормальному распределению, чем к равномерному.

Второй способ построения пробит-графиков отличается от описанного выше только способом построения эмпирической функции распределения. На рис. 9.5 показаны выборка объемом 19 значений, имеющая стандартное нормальное распределение, и вариационный ряд, построенный по этой выборке. (Вариационный ряд построен с помощью сортировки выборочных значений; команда Данные^Сортировка.) Поскольку вероятность того, что случайная величина X примет значение из интервала $(x_{(i-1)}, x_{(i)})$, образованного последовательными порядковыми статистиками $x_{(i-1)}$ и $x_{(i)}$, не зависит от распределения и всегда равна $1/(n + 1)$ (см. раздел 2.3.9), значение эмпирической функции распределения

где $\bar{\mu}_k = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^k$ ($k = 2, 3, 4$) — выборочные центральные моменты (о вычислении таких моментов речь идет в разделе 8.4). Если распределение выборки близко к нормальному, то выборочные среднеквадратические отклонения этих коэффициентов вычисляются соответственно по формулам

$$s_1 = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}} \text{ и } s_2 = \sqrt{\frac{24n(n-2)(n-3)}{(n+1)^2(n+3)(n+5)}}.$$

Иногда рекомендуется выборочные коэффициенты асимметрии и эксцесса вычислять по формулам

$$\bar{\beta}_1 = \frac{k_3}{\sqrt{k_2^3}}, \quad \bar{\beta}_2 = \frac{k_4}{k_2^2} - 3, \text{ где } k_2 = \frac{\bar{\mu}_2}{1 - \frac{1}{n}}, \quad k_3 = \frac{\bar{\mu}_3}{\left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right)},$$

$$k_4 = \frac{\bar{\mu}_4}{\left(1 - \frac{2}{n+1}\right)\left(1 - \frac{2}{n}\right)\left(1 - \frac{3}{n}\right)} - \frac{3\bar{\mu}_2^2}{\left(1 - \frac{2}{n}\right)\left(1 - \frac{3}{n}\right)}.$$

Если выборочное распределение нормально или близко к нормальному, то вычисленные по последним формулам $\bar{\beta}_1$ и $\bar{\beta}_2$ имеют асимптотически нормальные распределения с нулевыми математическими ожиданиями и среднеквадратическими отклонениями соответственно

$$s_1 = \sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \text{ и } s_2 = \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}}.$$

Считается, что если выполняется неравенство $|\bar{\beta}_1| \leq 3s_1$, то распределение симметрично. Если, кроме того, для коэффициента эксцесса выполняется неравенство $|\bar{\beta}_2| \leq 5s_2$, распределение можно считать нормальным. Реализовать приведенные формулы в Excel не представляет труда.

Если говорить о нормальном распределении, то для определения принадлежности выборочного распределения классу нормальных законов используется еще одна числовая характеристика — так называемое *нормированное среднее абсолютное отклонение*, определяемое формулой $\delta = \frac{M|X - MX|}{\sigma}$. Эта величина для

нормального распределения равна $\sqrt{2/\pi} = 0,79788$. Выборочное значение данного показателя вычисляется по формуле $\bar{\delta} = \frac{1}{ns} \sum_{i=1}^n |x_i - \bar{x}|$, где s — выборочное среднеквадратическое отклонение. Если выборочное распределение нормально или близко к нормальному, то распределение $\bar{\delta}$ асимптотически нормально с параметрами

$$M\bar{\delta} = \frac{2}{\sqrt{\pi(n-1)}} \frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} = \sqrt{\frac{2}{\pi}} \left(1 + \frac{2}{8n-9} + O\left(\frac{1}{n^2}\right)\right),$$

$$D\bar{\delta} = \frac{1}{n} \left\{ 1 + \frac{2}{\pi} \left[\sqrt{n(n-1)} + \arcsin \frac{1}{n-1} \right] \right\} - \frac{n-1}{\pi} \left[\frac{\Gamma\left(\frac{n+1}{2}\right)}{\Gamma\left(\frac{n}{2}\right)} \right]^2 = \frac{1}{n} \left(0,04507 - 0,0796 \frac{1}{n} + O\left(\frac{1}{n^2}\right) \right),$$

где $\Gamma(x)$ — гамма-функция Эйлера. Если выполняется неравенство $|\bar{\delta} - \sqrt{2/\pi}| \leq 0,7/\sqrt{n}$, то выборочное распределение можно считать нормальным или близким к нормальному.

Очевидно, что значения $\beta_1 = 0$, $\beta_2 = 0$ и $\delta = \sqrt{2/\pi}$ могут иметь распределения, отличные от нормального, и близость к этим значениям выборочных коэффициентов асимметрии, эксцесса и нормированного среднего абсолютного отклонения не гарантирует нормальности выборочного распределения. Описываемый ниже критерий, основанный на выборочных значениях этих показателей, служит, главным образом, не для проверки нормальности выборочного распределения, а для выявления отклонений выборочного распределения от нормального, точнее — для проверки гипотез $\beta_1 \neq 0$, $\beta_2 \neq 0$ и $\delta \neq \sqrt{2/\pi}$.

9.2.1. Критерии отклонения распределения от нормального

Статистическая модель. Выборка, состоящая из независимых выборочных значений x_1, x_2, \dots, x_n , получена из генеральной совокупности, имеющей нормальное распределение с неизвестными параметрами m и σ .

Для проверки значений коэффициентов β_1 , β_2 и δ можно сформулировать несколько гипотез, проверяя их значения поодиночке, попарно или совместно для всех трех коэффициентов. Покажем три критерия проверки гипотез о значениях этих коэффициентов по отдельности. Но, поскольку вычисления для всех трех критериев однотипны, описание их проведем параллельно, обозначая критерии как a , b и c .

Гипотезы

H_0 : а) $\beta_1 = 0$; б) $\beta_2 = 0$; в) $\delta = \sqrt{2/\pi}$

H_1 : а) $\beta_1 \neq 0$; б) $\beta_2 \neq 0$; в) $\delta \neq \sqrt{2/\pi}$

Задается уровень значимости α .

Вычисления

1. По выборочным значениям вычисляются первые четыре выборочных момента и выборочное среднее абсолютного отклонения по формулам

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2, \quad \bar{\mu}_3 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^3, \quad \bar{\mu}_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4, \\ d_n = \frac{1}{n} \sum_{i=1}^n |x_i - \bar{x}|.$$

2. Вычисляются критериальные статистики:

а) выборочный коэффициент асимметрии $\bar{\beta}_1 = \frac{\bar{\mu}_3}{\sqrt{s_n^3}}$;

б) выборочный коэффициент эксцесса $\bar{\beta}_2 = \frac{\bar{\mu}_4}{s_n^2} - 3$;

в) выборочное нормированное среднее абсолютного отклонения $\bar{\delta} = \frac{d_n}{s_n} - \sqrt{\frac{2}{\pi}}$.

3. Вычисляются выборочные среднеквадратические отклонения подсчитанных в предыдущем пункте величин

а) выборочного коэффициента асимметрии $s_1 = \sqrt{\frac{6(n-2)}{(n+1)(n+3)}}$;

б) выборочного коэффициента эксцесса $s_2 = \sqrt{\frac{24n(n-1)^2}{(n-3)(n-2)(n+3)(n+5)}}$;

в) выборочного нормированного среднего абсолютного отклонения $s_3 = \sqrt{\frac{1}{n} \left(0,04507 - 0,0796 \frac{1}{n} \right)}$.

Построение критической области. При условии истинности гипотезы H_0 все выборочные коэффициенты имеют асимптотически нормальные распределения с нулевыми математическими ожиданиями и соответствующими дисперсиями. Находится квантиль t порядка $1 - \alpha/2$ стандартного нормального распределения. Вычисляются критические значения:

для гипотезы а) — $t_{кр} = ts_1$;

для гипотезы б) — $t_{кр} = ts_2$;

для гипотезы в) — $t_{кр} = ts_3$.

Нулевая гипотеза принимается, если $|T| \leq t_{кр}$ (T — одна из подсчитанных в п. 2 критериальных статистик). В противном случае нулевая гипотеза отвергается.

Комментарии

1. Все три критерия являются приближенными. Их точность зависит от близости распределения критериальных статистик к нормальному для данного объема выборки n . Распределения выборочного коэффициента асимметрии и выборочного среднего абсолютного отклонения сходятся к нормальному достаточно быстро. Считается, что для этих величин достаточная точность достигается при $n > 50$. Распределение выборочного коэффициента эксцес-

са сходится к нормальному очень медленно — приемлемая точность достигается только для выборок, имеющих несколько тысяч значений. Для малых выборок существуют таблицы определения критических значений [4].

2. Если вычисляются все три критерия, гипотезу о нормальности распределения следует отвергнуть, когда хотя бы по одному критерию отвергается нулевая гипотеза.

Практическая реализация

На рис. 9.7 показан рабочий лист Excel, на котором реализованы все три описанных критерия. На этом же листе представлены формулы, по которым выполняются вычисления. (Не показаны формулы для вычисления количества выборочных значений, среднего, дисперсии и стандартного отклонения — они вычисляются по стандартным формулам.) В качестве тестовой выборки взята выборка из 100 значений, имеющих распределение Стьюдента со степенью свободы 2; значение степени свободы задается в ячейке G1 (о том, как моделировать распределение Стьюдента, речь идет в разделе 7.2). Выборочный коэффициент асимметрии назван Бета 1, выборочный коэффициент эксцесса — Бета 2, выборочное среднее абсолютного отклонения — Дельта.

E2		=НОРМСТОБР(1-E1/2)				
	A	B	C	D	E	F
1	Выборка	Количество	100	Альфа	0,05	Степень
2	-1,16376	Среднее	-0,28205	Квантиль	1,9599628	свободы
3	0,25829	Дисперсия	2,827384	Критическое бета 1	0,465969	=C11*E2
4	0,201651	Станд. отклонение	1,681483	Критическое бета 2	0,891287	=C12*E2
5	0,697704	Среднее абс. отклонений	1,120595	Критическое дельта	0,04124	=C13*E2
6	-5,59397	3-й момент	-1,42383	={СУММ((Выборка-C2)^3)/C1}		
7	0,340411	4-й момент	55,76392	={СУММ((Выборка-C2)^4)/C1}		
8	0,009791	Бета 1	-0,29949	=C6/СТЕПЕНЬ(C4;3)		
9	-0,99471	Бета 2	3,975636	=C7/(C3*C3)-3		
10	-2,62133	Дельта	-0,13145	=C5/C4-КОРЕНЬ(2/П())		
11	1,237888	Отклонение бета 1	0,237744	=КОРЕНЬ(6*(C1-2)/((C1+1)*(C1+3)))		
12	1,382357	Отклонение бета 2	0,454747	=КОРЕНЬ(24*C1*(C1-2)*(C1-3)/		
13	-0,30509	Отклонение дельта	0,021041	((C1+1)*(C1+1)*(C1+3)*(C1+5)))		
14	-0,07244					
15	0,109213	=СРОТКЛ(Выборка)		=КОРЕНЬ((0,04507-0,0796/C1)/C1)		
16	-1,083					
17	-2,09084					

Рис. 9.7. Формулы для критериев

Как видно на рис. 9.7, критерии по коэффициенту эксцесса и выборочному среднему абсолютного отклонения отклоняют гипотезу о нормальности распределения выборки. Если выборка будет иметь распределение Стьюдента со степенью свободы 6 (чтобы изменить выборку, достаточно изменить значение в ячейке G1), то, как показано на рис. 9.8, гипотезу о нормальности следует отвергнуть только по критерию среднего абсолютного отклонения. Таким образом, критерий по коэффициенту асимметрии в данном случае практически не работает, поскольку распределение Стьюдента симметрично, но другие критерии могут выявить отклонение от нормальности.

9.2.2. Критерий отклонения от распределения Пуассона

Если параметры гипотетического распределения связаны каким-либо соотношением (например, дисперсия распределения y при любом числе степеней свободы ровно в два раза больше математического ожидания; см. раздел 1.5.5), то выполнение этого соотношения для числовых характеристик данной выборки можно использовать как показатель того, что выборочное распределение совпадает с гипотетическим распределением. Но поскольку такое соотношение может иметь и распределение другого типа (как нулевые значения коэффициентов асимметрии и эксцесса в случае нормального распределения из предыдущего раздела), то чаще *невыполнение* этого соотношения используют как критерий отклонения от данного гипотетического распределения. На такой основе построен критерий отклонения от распределения Пуассона, математическое ожидание и дисперсии которого, как известно, совпадают (см. раздел 1.4.4).

	A	B	C	D	E	F	G
1	Выборка	Количество	100	Альфа	0,05	Степень	6
2	-0,0235	Среднее	0,108268	Квантиль	1,9599628	свободы	
3	1,50108	Дисперсия	1,53504	Критическое бета 1	0,465969	=C11*E2	
4	-0,21775	Станд. отклонение	1,238967	Критическое бета 2	0,891287	=C12*E2	
5	-0,58504	Среднее абс. отклонений	0,922155	Критическое дельта	0,04124	=C13*E2	
6	1,129414	3-й момент	-0,84642	{=СУММ((Выборка-C2)^3/C1)}			
7	1,18221	4-й момент	8,290297	{=СУММ((Выборка-C2)^4/C1)}			
8	-0,3409	Бета 1	-0,44505	=C6/СТЕПЕНЬ(C4;3)			
9	0,209093	Бета 2	0,518282	=C7/(C3^C3)-3			
10	3,348341	Дельта	-0,05359	=C5/C4-КОРЕНЬ(2/ПИ())			
11	0,480354	Отклонение бета 1	0,237744	=КОРЕНЬ(6*(C1-2)/((C1+1)*(C1+3)))			
12	-1,38971	Отклонение бета 2	0,454747	=КОРЕНЬ(24*C1*(C1-2)*(C1-3)/			
13	1,239782	Отклонение дельта	0,021041	((C1+1)*(C1+1)*(C1+3)*(C1+5)))			
14	1,764679						
15	1,949334	=СРОТКЛ(Выборка)		=КОРЕНЬ((0,04507-0,0796/C1)/C1)			
16	1,306428						
17	0,04648						

Рис. 9.8. Критерии для новой выборки

Статистическая модель. Выборка, состоящая из независимых выборочных значений x_1, x_2, \dots, x_n , является реализацией случайной величины X , имеющей распределение Пуассона с неизвестным параметром λ .

Гипотезы

H_0 : $DX = MX$

H_1 : $DX \neq MX$

Задается уровень значимости α .

Вычисления

1. По выборочным значениям вычисляются выборочные среднее и дисперсия

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, \quad s_n^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2.$$

2. Вычисляется критериальная статистика $T = \frac{(n-1)s_n^2}{\bar{x}}$.

Построение критической области. При условии истинности гипотезы H_0 критериальная статистика асимптотически имеет распределение χ^2 с $(n-2)$ степенью свободы. Находятся двухсторонние критические значения t_n и t_α как квантили соответственно порядка $\alpha/2$ и порядка $1 - \alpha/2$ распределения χ^2 с $(n-2)$ степенью свободы. Гипотеза H_0 принимается, если выполняется неравенство $t_n \leq T \leq t_\alpha$, иначе гипотеза H_0 отклоняется.

Комментарии

1. Критерий является приближенным. Он применяется, если $n\bar{x} \geq 10$.
2. По сути, это критерий проверки гипотезы о значении дисперсии нормального распределения (см. раздел 2.4.1), где точное значение дисперсии заменено выборочным средним. Отсюда требование, чтобы $n\bar{x}$ было достаточно большим, — тогда распределение Пуассона можно аппроксимировать нормальным распределением.

Практическая реализация в Excel этого критерия очевидна и не вызывает затруднений.

Далее рассмотрим критерий согласия χ^2 (критерий Пирсона) и критерий Колмогорова.

9.3. Критерий χ^2

Теоретическое описание этого критерия дано в разделе 2.4.3. Здесь приведем его практическую реализацию для двух случаев: для непрерывных распределений и для дискретных. Сначала рассмотрим случай дискретных распределений.

9.3.1. Критерий χ^2 для дискретных распределений

Статистическая модель. Выборка из дискретной генеральной совокупности представлена в виде частотной таблицы, в которой для каждого значения $x_1, x_2,$

..., x_k указываются соответствующие частоты f_1, f_2, \dots, f_k , при этом $\sum_{i=1}^k f_i = n$.

Функция распределения $F(u)$ генеральной совокупности зависит от m параметров, из которых m_1 параметров неизвестны. (Например, $F(u)$ — функция биномиального распределения с параметрами n и p , при этом n известно, а значение p неизвестно. Тогда $m = 2$, а $m_1 = 1$.)

Гипотезы

H_0 : выборочные значения получены из генеральной совокупности с функцией распределения $F(u)$, зависящей от m параметров, из которых m_1 параметров определяются по выборочным значениям².

H_1 : нулевая гипотеза неверна.

² Здесь не указывается, как задается распределение: с помощью функции распределения или функции вероятностей (см. раздел 1.2.1). Здесь говорится лишь о том, что распределение известно, способ его задания влияет только на дальнейшие вычисления.

Задается уровень значимости α .

Вычисления

1. В предположении, что справедлива гипотеза H_0 , вычисляются ожидаемые значения частот v_i для всех значений x_i . Если распределение задается с помощью функции вероятностей (см. раздел 1.2.1), то $v_i = np_i = nP(X = x_i)$. Если известна функция распределения $F(u) = P(X < u)$, то $v_i = n[F(x_i) - F(x_{i-1})]$. В последнем случае для $i = 1$ $v_1 = nF(x_1)$.

2. Вычисляется критериальная статистика $T = \sum_{i=1}^k \frac{(f_i - v_i)^2}{v_i}$.

Построение критической области. При условии истинности гипотезы H_0 статистика T асимптотически имеет распределение χ^2 с $(k - m_1 - 1)$ степенью свободы.

Находится критическое значение $t_{кр}$ — квантиль порядка $1 - \alpha$ распределения χ^2 с $(k - m_1 - 1)$ степенью свободы. Гипотеза H_0 принимается, если $T \leq t_{кр}$. В противном случае гипотеза H_0 отклоняется.

В статистической литературе приводятся специальные таблицы, по которым на основе значений T и известного числа степеней подсчитаны вероятности $\alpha_{кр} = P(X > T)$ (X — случайная величина, имеющая распределение χ^2). Тогда, если $\alpha_{кр}$ больше заданного уровня значимости α , гипотеза H_0 принимается. В противном случае отклоняется. Это же значение $\alpha_{кр}$ подсчитывает функция Excel **CHI2TEST** (см. раздел 4.8.4).

Комментарий. Существуют различные мнения о том, какими должны быть величины ожидаемых частот $v_i = np_i$ (см., например, [13]). “Среднее” мнение таково, что большинство v_i должно быть больше 5 и не более 20% этих значений может быть меньше 5. Если малых значений частот слишком много, то их можно объединить. Также отметим, что если объем выборки достаточно большой, например $n \geq 50$, и при этом $k \geq 10$, тогда вопрос о значениях ожидаемых частот снимается.

Практическая реализация. На рис. 9.9 приведен рабочий лист Excel, на котором показаны все этапы выполнения критерия χ^2 . В столбце A содержится выборка из 100 значений (полученная с помощью средства Генерация случайных чисел), имеющая распределение Пуассона с параметром $\lambda = 1$. Диапазону ячеек, содержащему выборочные значения, присвоено имя Выборка. В столбце B подсчитаны объем выборки (формула **=СЧЁТ(Выборка)**), среднее, т.е. оценка λ (формула **=СРЗНАЧ(Выборка)**), и количество различных значений в выборке (формула массива **{=СУММ(1/СЧЁТЕСЛИ(Выборка;Выборка))}**). В столбце C записаны все различные выборочные значения, для чего использована формула, приведенная в разделе 8.3.1 (в этом разделе подробно рассмотрен процесс создания частотных таблиц). Частоты (столбец D) подсчитываются с помощью функции **ЧАСТОТА**. Ожидаемые частоты вычисляются с помощью функции **ПУАССОН** (см. раздел 4.6.11). В ячейке E2 записана формула **=ПУАССОН(C2;\$B\$6;0)*\$B\$4**, которая затем скопирована в ячейки E3:E7.

Для вычисления критериальной статистики T можно применить формулу массива

$$\{=\text{СУММ}(((\text{Частоты}-\text{Ожидаемые_частоты})^2)/\text{Ожидаемые_частоты})\},$$

если интервалу D2:D7, содержащему значения частот, присвоить имя Частоты, а интервалу E2:E7, содержащему значения ожидаемых частот, — имя Ожидаемые_частоты. В столбце G задано значение уровня значимости, подсчитаны значение степени свободы (формула =B2-2) и $t_{кр}$, значение квантиля распределения χ^2 (формула =ХИ2ОБР(G2;G4); описание функции ХИ2ОБР приведено в разделе 7.7.8). Как видно, в данном случае выборочные данные не противоречат гипотезе, что их распределение является распределением Пуассона с параметром $X = 1,12$. В ячейке G8 вычислено критическое значение $\alpha_{кр}$ (формула =ХИ2ТЕСТ(Частоты;Ожидаемые_частоты)). Поскольку $\alpha_{кр}$, равное 0,78725, значительно больше уровня значимости 0,05, нулевая гипотеза принимается.

9.3.2. Критерий χ^2 для непрерывных распределений

Статистическая модель. Выборка, состоящая из независимых выборочных значений x_1, x_2, \dots, x_n , получена из генеральной совокупности, имеющей функцию распределения $F(u)$. Функция $F(u)$ зависит от m параметров, из которых m_i параметров неизвестно.

G8 =ХИ2ТЕСТ(D2:D7;E2:E7)						
	А	В	С	Д	Е	Е
	Выборка	К-во различных значений	Значения	Ожидаемые частоты	Критериальная статистика	Уровень значимости
1						
2	0	6	0	35 32,62797946	2,426297342	0,05
3	3	Объем	3	10 7,639993655		Ст. свободы
4	1	100	1	33 36,543337		4
5	0	Среднее	2	20 20,46426872	Критическое значение t	
6	0	1,12	5	1 0,479180402		7,814724703
7	0		4	1 2,139198223	Критическое значение альфа	
	0					0,787252774
9	0					
10	0					
11	1					
12	3					
13	0					
14	3					
	0					

Рис. 9.9. Критерий χ^2 для дискретного распределения

Гипотезы

H_0 : выборочные значения получены из генеральной совокупности с функцией распределения $F(u)$, зависящей от m параметров, из которых m_x параметров определяются по выборочным значениям.

H_1 : нулевая гипотеза неверна.

Задается уровень значимости α .

Вычисления

1. Область возможных выборочных значений разбивается на k непересекающихся интервалов $\Delta_1 = (x^{(1)}, x^{(2)})$, $\Delta_2 = (x^{(2)}, x^{(3)})$, ..., $\Delta_k = (x^{(k)}, x^{(k+1)})$. (Определение таких интервалов рассмотрено ниже.)

2. Подсчитывается, сколько выборочных значений попало в каждый интервал Δ_i . Получаем ряд частот n_1, n_2, \dots, n_k (при этом должно выполняться равенство $n_1 + n_2 + \dots + n_k = n$, где n — объем выборки).
3. В предположении, что справедлива гипотеза H_0 , по формуле $v_i = n[F(x^{(i+1)}) - F(x^{(i)})]$ вычисляются ожидаемые значения частот, т.е. количества попаданий выборочных значений в каждый из интервалов Δ_i , где $x^{(i)}$ и $x^{(i+1)}$ — границы интервала Δ_i .
4. Вычисляется критериальная статистика $T = \sum_{i=1}^k \frac{(n_i - v_i)^2}{v_i}$.

Построение критической области. При условии истинности гипотезы H_0 статистика T асимптотически имеет распределение χ^2 с $(k - m_1 - 1)$ степенью свободы.

Вычисляется критическое значение критерия $t_{кр}$ — квантиль порядка $1 - \alpha$ распределения χ^2 с $(k - m_1 - 1)$ степенью свободы (для нахождения квантиля можно использовать функцию ХИ2ОБР). Гипотеза H_0 принимается, если $T \leq t_{кр}$. В противном случае гипотеза H_0 отклоняется.

На основе значения T можно также вычислить вероятность $\alpha_{кр} = P(X > T)$ (X — случайная величина, имеющая распределение χ^2 с числом степеней свободы $k - m_1 - 1$). Тогда, если $\alpha_{кр}$ больше заданного уровня значимости α , гипотеза H_0 принимается. В противном случае она отклоняется. Значение $\alpha_{кр}$ подсчитывает функция Excel ХИ2ТЕСТ (см. раздел 4.8.4).

Комментарии

1. Разбиение области выборочных значений на интервалы $\Delta_k = (x^{(k)}, x^{(k+1)})$ можно выполнить многими способами. Вот два основных подхода. Интервалы Δ_k подбираются таким образом, чтобы все ожидаемые частоты v_k были равными (другими словами, чтобы были равны вероятности попадания выборочных значений в эти интервалы), либо интервалы Δ_k строятся равной длины. Первый подход имеет определенные преимущества, поскольку можно заранее задать значения v_k , например, равными 5 или 6. Однако в таком случае интервалы Δ_k имеют разные длины (кроме случая равномерного распределения) и при их построении могут возникнуть определенные сложности. На практике чаще используется второй подход, при котором интервалы Δ_k предполагаются равной длины. Здесь существует своя проблема определения количества таких интервалов. Практическое правило рекомендует первоначально выбирать длину интервалов равной примерно $0,4z$, где z — среднеквадратическое отклонение выборки. После вычисления ожидаемых частот количество интервалов (и, соответственно, длина интервалов) может быть изменено таким образом, чтобы рассчитанные величины ожидаемых частот были не меньше некоторой заранее заданной величины (в качестве такой величины чаще всего опять выступает “магическое” число 5). Повторим рекомендацию из комментария к этому критерию для дискретных распределений: большинство v_i должно быть не меньше 5 и не более 20% этих значений может быть меньше 5 (но обязательно не меньше 1).

2. Необходимо помнить, что критерий χ^2 является все-таки приближенным. Поэтому надо проявлять "бдительность" и осторожность, когда значение критериальной статистики T близко к критическому значению $t_{кр}$. Кроме того, этот критерий не учитывает порядок выборочных значений (критерий не почувствует неестественность выборки, если, например, все малые выборочные значения сосредоточены в начале выборки, а большие — в конце). Поэтому, если есть возможность, для проверки гипотезы о принадлежности распределения выборки заданному классу распределений следует применять более мощный и более чувствительный критерий Колмогорова (см. следующий раздел).

Практическая реализация

Как указывалось выше в комментариях, существуют два подхода к определению интервалов. При первом подходе интервалы строятся таким образом, чтобы ожидаемые частоты для всех интервалов были равными (случай равновероятных интервалов). При втором подходе все интервалы имеют равные длины. Покажем реализацию критерия с использованием этих двух подходов к определению и построению интервалов. Рассмотрим сначала первый подход.

На рис. 9.10 показан рабочий лист Excel со всеми формулами, необходимыми для реализации критерия. В столбце A содержится выборка объемом 100 значений, имеющая стандартное нормальное распределение (получена с помощью средства Генерация случайных чисел). Диапазону ячеек, содержащему выборочные значения, присвоено имя **Выборка**. В столбце B с помощью стандартных формул подсчитаны основные характеристики выборки: среднее, стандартное отклонение, минимальное и максимальное значения, количество выборочных значений. Задаем величину ожидаемой частоты. Пусть это значение равно 5 (ячейка B12). Подсчитываем ожидаемую частоту (формула $=B12/B10$ в ячейке B14) и количество интервалов (формула $=B10/B12$ в ячейке B16). Пусть нулевая гипотеза состоит в том, что генеральная совокупность имеет стандартное нормальное распределение. (Обращаем внимание на то, что здесь параметры гипотетического (стандартного нормального) распределения не определяются на основе выборочных значений, поэтому количество степеней свободы будет на единицу меньше количества интервалов.)

Далее необходимо определить границы интервалов. Это можно сделать с помощью одной формулы массива, использующей функции **СТРОКА** и **ДВССЫЛ** (как в разделе 8.3.2 для определения интервалов при создании гистограмм). Здесь (для разнообразия) используем простые и "прозрачные" формулы, но за эту простоту заплатим дополнительным столбцом значений — в столбце C введены номера интервалов от 1 до 20. Тогда границу первого интервала (ячейка D2) вычисляет формула $=НОРМСТОБР(C2*\$B\$14)$, которая затем распространяется вниз до ячейки D20. Обращаем внимание, что для интервала 20 эта формула не используется; для этого интервала будет подсчитано число выборочных значений, которые превышают верхнюю границу 19-го интервала. Для вычисления частот применяется формула массива $\{=ЧАСТОТА(Выборка;Интервалы)\}$ (здесь диапазон ячеек D2:D20 назван **Интервалы**). Подчеркнем, что при вводе этой формулы необходимо выделить диапазон E2:E21, а не E2:E20.

Поскольку в данном случае формально нет массива ожидаемых частот, функция **ХИ2ТЕСТ** не применима. Поэтому вычислим критериальную статистику T

и критическое значение $t_{кр}$, по которым будем судить о значимости нулевой гипотезы. Статистика T вычислена в ячейке F2 с использованием формулы массива

$$\{=СУММ(((Частота-\$B\$12)*2)/\$B\$12)\}.$$

Здесь диапазону частот E2:E21 присвоено имя Частота и в ячейке B12 содержится заданное значение ожидаемых частот. В ячейке F4 задано значение уровня значимости, в ячейке F6 вычислено количество степеней свободы (на единицу меньше количества интервалов). Критическое значение ξ_{α} записано в ячейке F8; оно вычисляется по формуле $=ХН20ВР(F4;F6)$. Сравнив значения в ячейках F2 и F8, приходим к выводу, что при заданном уровне значимости нет оснований отвергать нулевую гипотезу. Критическое значение вероятности α^* , которое обычно вычисляет функция ХИ2ТЕСТ, можно вычислить по формуле $=ХИ/12РАСн(F2;F6)$ (ячейка F10). Это значение также показывает, что следует принять нулевую гипотезу.

F2		{=СУММ(((Частота-\$B\$12)*2)/\$B\$12)}				
	A	B	C	D	E	F
1	Выборка	Среднее	№ интервала	Интервалы	Частота	Критериальная статистика
2	-0,30023	0,090500441	1	-1,6448535	6	11,2
3	1,27647	Ст. отклонение	2	-1,2815519	4	Уровень значимости
4	-2,18359	1,000184356	3	-1,0364335	4	0,05
5	-1,0867	Мин. значение	4	-0,841621	6	Ст. свободы
6	-1,84691	-2,191118256	5	-0,6744895	4	19
7	-2,11793	Макс. значение	6	-0,5244005	3	Критическое значение t
8	0,13485	2,343476808	7	-0,3853206	2	30,14351
9	-0,37024	Количество	8	-0,2533472	4	Критическое альфа
10	-0,18616	100	9	-0,1256612	6	0,916927
11	0,88567	Ожидаемая частота	10	5,4714E-10	3	
12	1,66146	5	11	0,12566125	7	
13	0,90219	Ожидаемая частота	12	0,25334724	5	
14	-0,5238	0,05	13	0,3853206	3	
15	0,75761	К-во интервалов	14	0,52440046	6	
16	-1,52157	20	15	0,67448953	5	
17	0,02812		16	0,84162104	6	
18	-1,74248		17	1,03643347	9	
19	1,44767		18	1,28155194	6	
20	0,75771		19	1,64485348	7	
21	0,59574		20		4	
22	0,69399					

Рис. 9.10. Критерий χ^2 для равновероятных интервалов

Теперь рассмотрим данный критерий для случая равных интервалов. На рис. 9.11 представлен рабочий лист, содержащий ту же выборку, что и на рис. 9.10. Длину интервала выбираем равной 0,4 стандартного отклонения, т.е. равной 0,4 (ячейка B12). Далее определяем нижнюю границу интервалов; она должна быть больше минимального выборочного значения. Здесь эта нижняя граница выбрана равной -2 (ячейка B14). Верхняя граница интервалов должна быть меньше максимального значения; принимаем верхнюю границу равной 2 (ячейка B16). Вычисляем количество интервалов (формула $=(B16-B14)/B12$ в ячейке B18). Значения нижней и верхней границ выбирают таким образом, чтобы вычисленное количество интервалов было целым числом. Отметим, что общее число интервалов будет не 10, а 12, поскольку

имеются еще два интервала: один, содержащий значения, меньшие нижней границы, и второй, содержащий значения, большие верхней границы. В столбце D вычисляются границы интервалов. В ячейке D2 записана формула $=\$B\$14+(C2-1)*\$B\12 , которая затем копируется в диапазон D3:D12. Диапазону D2:D12 присвоено имя Границы.

Теперь подсчитываются значения выборочных частот (столбец E, формула массива $\{=\text{ЧАСТОТА}(\text{Выборка};\text{Границы})\}$) и ожидаемых частот (столбец F). Для вычисления ожидаемых частот используются такие формулы: в ячейке F2 — $=\text{НОРМСТРАСП}(D2)*\$B\10 , в ячейке F3 — $=(\text{НОРМСТРАСП}(D3)-\text{НОРМСТРАСП}(D2))*\$B\$10$, которая копируется в диапазон F4:F12. В ячейке F13 записана формула $=(1-\text{НОРМСТРАСП}(D12))*\$B\$10$.

Далее вычисляется значение критериальной статистики T по формуле массива (ячейка G2)

$$\{=\text{СУММ}(((\text{Частота}-\text{Ожидаемые_частоты})^2)/\text{Ожидаемые_частоты})\}.$$

В ячейке G8 вычисляется критическое значение $t_{кр}$ по формуле $=\text{ХИ2ОБР}(G4;G6)$, а в ячейке G10 — критическое значение вероятности $\alpha_{кр}$ по формуле $=\text{ХИ2ТЕСТ}(\text{Частота};\text{Ожидаемые_частоты})$. Сравнение вычисленных значений T и $t_{кр}$, а также значения уровня значимости со значением $\alpha_{кр}$ показывают, что выборочные значения не противоречат нулевой гипотезе. Отметим также, что значение $\alpha_{кр}$ здесь намного меньше, чем в предыдущем примере. Это говорит о том, что данный критерий в случае равновероятных интервалов более точен, чем в случае равных интервалов.

G2		{=СУММ(((Частота-Ожидаемые_частоты)^2)/Ожидаемые_частоты)}						
	A	B	C	D	E	F	G	H
1	Выборка	Среднее	Мз	Границы	Частота	Ожидаемые частоты	Критериальная статистика	
2	-0,300232	0,090500441	1	-2	3	2,275006204	8,86367	
3	1,276474	Ст. отклонение	2	-1,6	3	3,204822742	Уровень значимости	
4	-2,193588	1,000184358	3	-1,2	4	6,027044226	0,05	
5	-1,088701	Мин. значение	4	-0,8	12	9,678580222	Ст. свободы	
6	-1,846911	-2,191118256	5	-0,4	7	13,27228695	11	
7	-2,117931	Макс. значение	6	0	13	15,54216964	Критическое значение t	
8	0,134853	2,343478808	7	0,4	15	15,54216968	19,6752	
9	-0,370241	Количество	8	0,8	17	13,27228695	Критическое альфа	
10	-0,186158	100	9	1,2	12	9,678580222	0,63447	
11	0,885873	Длина интервала	10	1,6	9	8,027044226		
12	1,681458	0,4	11	2	4	3,204822742		
13	0,902191	Нижняя граница	12		1	2,275006204		
14	-0,523795	-2						
15	0,757611	Верхняя граница						
16	-1,521571	2						
17	0,028117	К-во интервалов						
18	-1,742483	10						
19	1,44787							
20	0,757714							

Рис. 9.11. Критерий χ^2 для равных интервалов

9.4. Критерий Колмогорова

Данный критерий более мощный, чем критерий χ^2 . Он предполагает непрерывность распределений. Однако на практике критерий часто используется для сгруппированных данных (т.е. данных, представленных в виде частотной таблицы) и даже для дискретных распределений. Общее описание этого критерия дано в разделе 2.4.3.

Статистическая модель. Выборка, состоящая из независимых выборочных значений x_1, x_2, \dots, x_n , получена из генеральной совокупности, распределение которой предполагается непрерывным.

Гипотезы

H_0 : выборочные значения получены из генеральной совокупности с функцией распределения $F(u)$.

H_1 : нулевая гипотеза неверна.

Задается уровень значимости α .

Вычисления

1. По выборке x_1, x_2, \dots, x_n строится вариационный ряд $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$.

2. Вычисляются *кумулятивные разности*: $D_m^+ = \frac{m}{n} - F(x_{(m)})$ и $D_m^- = F(x_{(m)}) - \frac{m-1}{n}$,
 $m = 1, 2, \dots, n$.

3. Вычисляется критериальная статистика $D_n = \max_{1 \leq m \leq n} (D_m^+, D_m^-)$.

Построение критической области. При условии истинности гипотезы H_0 статистика D_n имеет так называемое распределение Колмогорова–Смирнова.

Находится критическое значение $t_{кр}$ — квантиль порядка $1 - \alpha$ распределения Колмогорова–Смирнова. Гипотеза H_0 принимается, если $D_n \leq t_{кр}$. В противном случае гипотеза H_0 отклоняется.

Комментарии

1. Для нахождения квантилей распределения Колмогорова–Смирнова существуют специальные таблицы, которые приведены во многих книгах по математической статистике. При $n \geq 10$ и $0,01 \leq \alpha \leq 0,2$ можно воспользоваться

приближенной формулой для вычисления $t_{кр}$: $t_{кр} \approx \sqrt{\frac{-\ln(0,5\alpha)}{2n}} - \frac{1}{6n}$ [4].

2. В литературе по математической статистике можно встретить упрощенный подход к вычислению критериальной статистики D_n , которую рекомендуют вычислять либо как $D_n = \max_{1 \leq m \leq n} (D_m^+)$, либо как $D_n = \max_{1 \leq m \leq n} (D_m^-)$. Как указано в [4], это неправильный подход, который может привести к неверным результатам, особенно при малых объемах выборки.

3. В этом критерии предполагается, что гипотетическое распределение известно точно. Если же параметры данного распределения определяются на основе выборочных значений, то необходима осторожность в применении критерия, особенно в случае, когда значение критериальной статистики близко к критическому значению.

Практическая реализация

Для иллюстрации применения описываемого критерия используем ту же выборку, что и в предыдущем разделе. Здесь также проверим гипотезу, что распределение генеральной совокупности имеет стандартное нормальное распределение. Выборка, имеющая стандартное нормальное распределение, показана на рис. 9.12 в столбце А. Диапазону ячеек, содержащему выборочные значения, присвоено имя Выборка. Этот диапазон скопирован в столбец В, в котором проведена его сортировка в порядке возрастания. В результате получен вариационный ряд. В столбце С записаны ранги выборочных значений. Поскольку для непрерывных распределений вероятность появления одинаковых значений в выборке равна нулю, ранги в данном случае просто совпадают с номерами порядковых статистик. Поэтому их можно не вычислять, а вводить как члены арифметической прогрессии с шагом 1 и начальным значением 1 (такая прогрессия вводится с помощью команды ПравкаЗаполнитьПрогрессия).

Далее вычисляются кумулятивные разности D_n^+ и D_n^- (записаны в столбце D и E; соответствующим диапазонам присвоены имена D_плюс и D_минус). Для вычисления D_n^+ в ячейке D2 введена формула $=C2/СЧЕТ(Выборка)-НОРМСТРАСП(B2)$, которая затем скопирована в остальные ячейки диапазона D_плюс. Аналогично для вычисления D_n^- в ячейку E2 записана формула $=НОРМСТРАСП(B2)-(C2-1)/СЧЕТ(Выборка)$, которая копируется вниз на весь диапазон D_минус.

F6 * =КОРЕНЬ(-LN(0.5*F4Y/(2*СЧЕТ(Выборка)))-1/(6*СЧЕТ(Выборка)))							
	A	B	C	D	E	F	G
1	Выборка	Вариационный ряд	Ранги	D-плюс	D-минус	Dn	
2	-0,30023	-2,191118256	1	-0,00422	0,01422	0,091559835	
3	1,276474	-2,18358764	2	0,005504	0,0045	Уровень значимости	
4	-2,18359	-2,117931217	3	0,01291	-0,00291		0,05
5	-1,0867	-1,846910891	4	0,00762	0,00238	Критическое значение	
6	-1,84691	-1,760936357	5	0,010875	-0,00088		0,134143485
7	-2,11793	-1,742482709	6	0,019288	-0,00929		
8	0,134853	-1,53858764	7	0,008047	0,00195		
9	-0,37024	-1,521570994	8	0,015942	-0,00594		
10	-0,18616	-1,396169864	9	0,008668	0,00133		
11	0,865673	-1,308389983	10	0,004629	0,00537		
12	1,661456	-1,167247774	11	-0,01156	0,02156		
13	0,902191	-1,116736712	12	-0,01205	0,02205		
14	-0,5238	-1,086700649	13	-0,00858	0,01858		
15	0,757611	-1,054675067	14	-0,00579	0,01579		
16	-1,52157	-1,029657142	15	-0,00159	0,01159		
17	0,028117	-1,026930931	16	0,007773	0,00223		
18	-1,74248	-0,984935014	17	0,007672	0,00233		
19	1,44767	-0,981090125	18	0,016726	-0,00673		
20	0,757714	-0,941504368	19	0,016777	-0,00678		
21	0,595742	-0,898178314	20	0,015455	-0,00545		

Рис. 9.12. Реализация критерия Колмогорова

Значение критериальной статистики D_n вычисляется в ячейке F2 по формуле $=\text{МАКС}(D_плюс; D_минус)$, а критическое значение $t_{кр}$ — в ячейке F2 по формуле $=\text{КОРЕНЬ}(-\text{LN}(0,5 \cdot F4) / (2 \cdot \text{СЧЁТ}(\text{Выборка}))) - 1 / (6 \cdot \text{СЧЁТ}(\text{Выборка}))$.

Как видно из результатов расчета, при заданном уровне значимости следует принять гипотезу о стандартном нормальном распределении генеральной совокупности.

Интервальное оценивание параметров распределения

В главе 8 в качестве одного из этапов предварительного анализа описано вычисление точечных оценок параметров выборочного распределения. Но для "полноценной" оценки неизвестных параметров только точечных оценок недостаточно — необходима какая-нибудь мера точности этих оценок. Как указывалось в главе 2, такой мерой точности могут служить доверительные интервалы. В разделе 2.2 главы 2 даны общие определения, относящиеся к построению доверительных интервалов. В данной главе рассмотрим конкретные методы построения таких интервалов.

Наиболее точные доверительные интервалы строятся на основе априорных предположений о классе распределений, которому, возможно, принадлежит распределение данной выборки. (Такие предположения должны подтверждаться на основе критериев проверки гипотез о распределениях, описанных в главе 9.) Доверительные интервалы, построенные без предположений о типе распределения выборки (или с минимальными предположениями, например с предположением только о симметричности распределения), как правило, основаны на асимптотических свойствах выборочных статистик и имеют приемлемую точность только для достаточно больших выборок. Ниже приведем несколько способов построения таких интервалов. Но большинство методов построения доверительных интервалов все-таки разработано для конкретных распределений.

Далее в этой главе предполагается, что необходимые точечные оценки параметров уже подсчитаны (см. раздел 8.4), за исключением оценок для некоторых конкретных распределений. Такие оценки (в частности, для параметров равномерного распределения) будут показаны в этой главе.

10.1. Общие доверительные интервалы для математического ожидания

Общие положения, на основе которых построены описываемые ниже методы, приведены в разделе 2.3.1.

10.1.1. Общая модель при известной дисперсии

Статистическая модель. Произвольное распределение генеральной совокупности с конечной известной дисперсией σ^2 .

Доверительный интервал для математического ожидания строится следующим образом.

1. Задается доверительный уровень α .

2. Из равенства $\alpha = 1 - 1/k^2$ определяется значение k : $k = \frac{1}{\sqrt{1-\alpha}}$.

3. Вычисляется доверительный интервал: $\left(\bar{x} - k \frac{\sigma}{\sqrt{n}}, \bar{x} + k \frac{\sigma}{\sqrt{n}} \right)$.

Комментарии

1. В рамках такой модели доверительный интервал для неизвестного математического ожидания можно построить только на основании неравенства Чебышева (см. раздел 1.2.4), которое в данном случае будет иметь вид

$$P(|\bar{x} - MX| \leq k \frac{\sigma}{\sqrt{n}}) \leq 1 - \frac{1}{k^2}.$$

2. В таком случае не рекомендуется брать большое значение α , поскольку это значительно снижает точность интервальной оценки.

Практическая реализация в Excel здесь тривиальна и поэтому не приводится.

10.1.2. Одномодальное симметричное распределение при известной дисперсии

Статистическая модель. Генеральная совокупность имеет симметричное одномодальное распределение с известной конечной дисперсией σ^2 .

Доверительный интервал строится следующим образом.

1. Задается доверительный уровень α .

2. Определяется значение k : $k = \frac{3}{2\sqrt{1-\alpha}}$.

3. Вычисляется доверительный интервал: $\left(\bar{x} - k \frac{\sigma}{\sqrt{n}}, \bar{x} + k \frac{\sigma}{\sqrt{n}} \right)$.

Комментарии

1. В этой статистической модели распределение статистики \bar{x} также будет симметричным и одномодальным. Поэтому для построения интервальных оценок можно воспользоваться неравенством Гаусса, которое в данном

случае будет иметь вид $P(|\bar{x} - MX| \leq k \frac{\sigma}{\sqrt{n}}) \leq 1 - \frac{4}{9k^2}$.

2. В этой модели существенно условие симметричности распределения, от которого нельзя освободиться (см. раздел 2.3.1).

Практическая реализация в Excel тривиальна.

10.1.3. Общая модель с неизвестной дисперсией

Статистическая модель. Произвольное распределение генеральной совокупности с конечным четвертым моментом и неизвестной дисперсией. Объем выборки n больше 30.

Доверительный интервал в данной статистической модели строится следующим образом.

1. Вычисляются точечные оценки \bar{x} и $S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$.
2. Задается доверительный уровень α .
3. Из уравнения $\alpha = 2F_{n-1}(k) - 1$, где F_{n-1} — функция распределения Стюдента с $(n - 1)$ степенью свободы, вычисляется значение k : $k = F_{n-1}^{-1}\left(\frac{1+\alpha}{2}\right)$,
 F_{n-1}^{-1} — функция, обратная к функции распределения Стюдента.
4. Вычисляется доверительный интервал: $\left(\bar{x} - k \frac{S_n}{\sqrt{n-1}}, \bar{x} + k \frac{S_n}{\sqrt{n-1}}\right)$.

Комментарии

1. В данной модели интервальные оценки построены на основе асимптотических свойств статистики \bar{x} (см. раздел 2.2).
2. В принципе, это тот же метод построения доверительного интервала для математического ожидания нормальной совокупности на основе распределения Стюдента (см. раздел 10.3.2), но здесь в статистической модели требуется достаточно большой объем выборки.

Практическая реализация

Практическое построение этого доверительного интервала в Excel не вызывает особых затруднений. На рис. 10.1 показан рабочий лист, содержащий все необходимые формулы для вычисления доверительного интервала. В столбце A содержится выборка, равномерно распределенная на интервале $[0, 10]$. (Эта выборка создана средством Генерация случайных чисел.) Диапазон ячеек, содержащий выборочные значения, назван Выборка.

D4		k = СТЬЮДРАСПОБР((1-D2)/2,B2-1)			
	A	B	C	D	E
1	Выборка	Объем выборки		Доверительный уровень	
2	3,8200018	100		0,95	
3	1,0068056	Среднее		Коэффициент k	
4	5,9648427	4,854570147		2,2760014	
5	8,9910581	Ст. отклонение		Граница	
6	8,8460952	3,092356913		0,70736658	
7	9,5846431				
8	0,1449629	Доверительный интервал			
9	4,074221	4,147203562		5,56193673	
10	8,6324656				
11	1,3858455	Длина интервала			
12	2,4503311	1,414733169			
13	0,4547258				
14	0,3238014				
15	1,6412854				
16	2,1961119				
17	0,1709037				
18	2,8504288				

Рис. 10.1. Построение доверительного интервала

Для выборки подсчитаны количество значений (ячейка B2, формула =СЧЁТ(Выборка)), выборочное среднее (ячейка B4, формула =СРЗНАЧ(Выборка)) и выборочное среднеквадратическое отклонение (ячейка B6, формула =СТАНДОТКЛОН(Выборка)). В столбце D задан доверительный уровень (ячейка D2), в ячейке D4 подсчитан коэффициент k по формуле

$$=СТЮДРАСПОБР((1-D2)/2;B2-1).$$

В ячейке D6 вычисляется величина $k \frac{S_n}{\sqrt{n-1}}$ по формуле =D4*B6/КОРЕНЬ(B2-1).

Наконец, вычисляются границы доверительного интервала: нижняя граница — по формуле =B4-D6 (ячейка B9) и верхняя — по формуле =B4+D6 (ячейка D9). Конечно, можно избежать промежуточных вычислений, выполненных в ячейках D4 и D6, и найти границы доверительного интервала с помощью одной формулы. Однако эти дополнительные вычисленные значения могут использоваться в анализе полученного результата.

10.2. Общий доверительный интервал для дисперсии

Если нет априорных предположений о типе распределения генеральной совокупности, то единственным способом построения доверительного интервала для неизвестной дисперсии является использование асимптотической нормальности распределения статистик для вычисления моментов генеральной совокупности (см. раздел 2.3.2).

Статистическая модель. Произвольное распределение генеральной совокупности с конечным четвертым моментом. Объем выборки — не менее 50.

Доверительный интервал в данной статистической модели строится следующим образом.

1. Вычисляются точечные оценки \bar{x} , S_n^2 и 4-го центрального момента

$$\bar{\mu}_4 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^4.$$

2. Вычисляется оценка среднеквадратического отклонения статистики S_n^2 по

$$\text{формуле } \sigma(S_n^2) = \sqrt{\frac{\bar{\mu}_4 - S_n^2}{n}}.$$

3. Задается доверительный уровень α .

4. Из уравнения $\alpha = 2\Phi(k) - 1$, где Φ — функция распределения стандартного нормального закона, вычисляется значение k : $k = \Phi^{-1}\left(\frac{1+\alpha}{2}\right)$, Φ^{-1} — функция, обратная к функции распределения стандартного нормального закона.

5. Вычисляется доверительный интервал: $(S_n^2 - k\sigma(S_n^2), S_n^2 + k\sigma(S_n^2))$.

Комментарий. Этот способ построения доверительного интервала является приближенным и дает удовлетворительные результаты только для достаточно больших выборок.

Практическая реализация

На рис. 10.2 показан рабочий лист Excel, на котором построен доверительный интервал для дисперсии выборки, равномерно распределенной на интервале [0, 10]. Эта выборка, записанная в столбце А, создана средством Генерация случайных чисел. Диапазон ячеек, содержащий выборочные значения, назван Выборка.

B10		A: {=СУММ(((Выборка-B4)^4)/B2)}						
	A	B	C	D	E	F	G	H
1	Выборка	Объем выборки	Доверительный уровень					
2	3,8200018	100	0,95					
3	1,0068056	Среднее	Коэффициент k					
4	5,9648427	4,854570147	1,959963					
5	8,9910581	Ст. отклонение	Граница					
6	8,8460952	3,092356913	1,524787					
7	9,5846431	Дисперсия	Доверительный интервал					
8	0,1449629	9,467044562	Доверительный интервал					
	4,074221	4-й момент	7,942258 10,99183					
9	8,6324656	150,14825						
1	1,3858455	Ср. отклонение дисперсии	Истинное значение дисперсии					
2	2,4503311	0,777967334	8,333333					
	0,4547258							
	0,3238014							
5	1,6412854							
6	2,1961119							

Рис. 10.2. Построение доверительного интервала для дисперсии

В столбце В подсчитаны необходимые статистические характеристики выборки: объем выборки, среднее, стандартное отклонение, выборочная дисперсия (ячейка B8, формула =ДИСПР(Выборка)), выборочный 4-й центральный момент (ячейка B10, формула массива {=СУММ((Выборка-B4)^4/B2)}) и среднеквадратическое отклонение дисперсии (ячейка B12, формула =КОРЕНЬ((B10-B8*B8)/B2)). В столбце D записано значение доверительного уровня (ячейка D2), подсчитаны значение коэффициента k (ячейка D4, формула =НОРМСТОБР((1+D2)/2)) и значение величины $k\sigma(S_n^2)$ (ячейка D6, формула =D4*B12).

После проведенных вычислений границы доверительного интервала (ячейки D9 и E9) вычисляются по простым формулам: =B8-D6 — для нижней границы и =B8+D6 — для верхней границы. В ячейке E12 приведено истинное значение дисперсии. Как видно, точечная оценка дисперсии значительно далека от истинного значения дисперсии, но доверительный интервал покрывает это значение даже с вероятностью 0,9 (рис. 10.3).

Отметим, что при необходимости любую границу доверительного интервала можно вычислить с помощью одной формулы Excel без показанных здесь промежуточных вычислений.

Другие интервальные оценки для дисперсий конкретных распределений будут показаны ниже.

	A	B	C	D	E	F	G
1	Выборка	Объем выборки		Доверительный уровень			
2	3,8200018	100		0,9			
3	1,0068056	Среднее		Коэффициент k			
4	5,9648427	4,854570147		1,644853			
5	8,9910581	Ст. отклонение		Граница			
6	8,8460952	3,092356913		1,279642			
7	9,5846431	Дисперсия					
8	0,1449629	9,467044562		Доверительный интервал			
9	4,074221	4-й момент		8,187402	10,74669		
10	8,6324656	150,14825					
11	1,3858455	Ср. отклонение дисперсии		Истинное значение дисперсии			
12	2,4503311	0,777967334		8,333333			
13	0,4547258						
14	0,3238014						
15	1,6412854						
16	2,1961119						

Рис. 10.3. Доверительный интервал с доверительным уровнем 0,9

10.3. Интервальные оценки параметров нормального распределения

Общие теоретические положения, на основе которых строятся описываемые ниже доверительные интервалы, приведены в разделе 2.3.6.

Статистическая модель. Генеральная совокупность имеет нормальное распределение с математическим ожиданием m и дисперсией σ^2 .

10.3.1. Интервальные оценки для неизвестного математического ожидания при известной дисперсии

Предполагается, что математическое ожидание m распределения генеральной совокупности неизвестно, но известна ее дисперсия σ . Доверительный интервал для m строится следующим образом.

1. Вычисляется точечная оценка $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

2. Задается доверительный уровень α .

3. Из уравнения $\alpha = 2\Phi(k) - 1$, где Φ — функция распределения стандартного нормального закона, вычисляется значение k : $k = \Phi^{-1}\left(\frac{1+\alpha}{2}\right)$, Φ^{-1} — функция, обратная к функции распределения стандартного нормального закона.

4. Вычисляется доверительный интервал: $\left(\bar{x} - k \frac{\sigma}{\sqrt{n}}, \bar{x} + k \frac{\sigma}{\sqrt{n}}\right)$.

Комментарии

1. Этот метод устойчив при умеренных отклонениях от нормальности.

2. Поскольку распределение выборочного среднего асимптотически нормально, этот метод можно применять для любых выборок, если их объем достаточно большой (по крайней мере, больше 30) и известна дисперсия.

Практическая реализация

Реализация этого метода построения доверительного интервала с соответствующими формулами показана на рис. 10.4. В данном примере выборка имеет нормальное распределение с параметрами $m = -1$ и дисперсией $\sigma^2 = 4$. Отметим, что в Excel для вычисления значения $k \frac{\sigma}{\sqrt{n}}$ предусмотрена функция ДОВЕРИТ (см. раздел 4.11.2), которая использована здесь в ячейке C10.

	A	B	C	D	E	F	G
1	Выборка	Известное значение дисперсии					
2	-1,60046	4					
3	-3,55537	Доверительный уровень					
4	-0,51149	0,95					
5	1,552947	Объем выборки					
6	1,3967	50 = СЧЕТ(Выборка)					
7	2,466266	Среднее					
8	-5,36718	-1,21401 = СРЗНАЧ(Выборка)					
9	-1,46836	Граница					
10	1,190045	0,554361 = ДОВЕРИТ(1-C4;КОРЕНЬ(C2);C6)					
11	-3,1734						
12	-2,38041	Доверительный интервал					
13	-4,38086	-1,76837 -0,65965					
14	-4,69382	=C8-C10 =C8+C10					
15	-2,95526						
16	-2,54701						
17	-5,23586						
18	-2,13585						

Рис. 10.4. Построение доверительного интервала для математического ожидания при известной дисперсии

10.3.2. Интервальные оценки для неизвестного математического ожидания при неизвестной дисперсии

Здесь предполагается, что математическое ожидание m и дисперсия σ^2 распределения генеральной совокупности неизвестны. Доверительный интервал для m строится следующим образом.

1. Вычисляются точечные оценки $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ и $S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$.
2. Задается доверительный уровень α .

3. Из уравнения $\alpha = 2F_{n-1}(k) - 1$, где F_{n-1} — функция распределения Стьюдента с $(n - 1)$ степенью свободы, вычисляется значение k : $k = F_{n-1}^{-1}\left(\frac{1+\alpha}{2}\right)$.
 F_{n-1}^{-1} — функция, обратная к функции распределения Стьюдента.

4. Вычисляется доверительный интервал: $\left(\bar{x} - k \frac{S_n}{\sqrt{n-1}}, \bar{x} + k \frac{S_n}{\sqrt{n-1}}\right)$.

Комментарии

1. Метод устойчив при умеренных отклонениях от нормальности.
2. Поскольку распределение выборочного среднего асимптотически нормально, этот метод можно применять для любых распределений, если объем выборки достаточно большой (по крайней мере, больше 30) и математическое ожидание и дисперсия распределения независимы.
3. Если дисперсия известна или оценивается на основании каких-либо иных данных, кроме выборочных значений, то следует применять метод, описанный в предыдущем разделе.

Практическая реализация

Для иллюстрации метода используем ту же выборку, что и в предыдущем примере (она имеет нормальное распределение с параметрами $m = -1$ и дисперсией $\sigma^2 = 4$). Реализация метода с соответствующими формулами показана на рис. 10.5. Отметим, что для вычисления коэффициента k используется функция СТЬЮДРАСПОБР (см. раздел 4.7.7).

B6		=СТАНДОТКЛОН(Выборка)					
	A	B	C	D	E	F	G
1	Выборка	Объем выборки		Доверительный уровень			
2	-1,60046	50		0,9			
3	-3,55537	Среднее		Коэффициент k			
4	-0,51149	-1,214006513		2,009574	=СТЬЮДРАСПОБР((1-D2)/2,B2-1)		
5	1,552947	Ст. отклонение		Граница			
6	1,3967	2,326813075		0,6679862	=D4*B6/КОРЕНЬ(B2-1)		
7	2,466266	=СТАНДОТКЛОН(Выборка)					
8	-5,36718			Доверительный интервал			
9	-1,46836			-1,8819927	-0,5460204		
10	1,190045			=B4-D6	=B4+D6		
11	-3,1734						
12	-2,38041						
13	-4,38086						
14	-4,89382						
15	-2,95526						
16	-2,54701						
17	-5,23586						
18	-2,13585						

Рис. 10.5. Построение доверительного интервала для математического ожидания при известной дисперсии

10.3.3. Интервальные оценки для неизвестной дисперсии при известном математическом ожидании

Предполагается, что математическое ожидание m распределения генеральной совокупности известно, но неизвестна ее дисперсия σ^2 . Доверительный интервал для σ^2 строится следующим образом.

1. Вычисляется выборочное среднее $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$; точечная оценка дисперсии вычисляется по формуле $S_n^2 = \frac{1}{n} \sum_{i=1}^n x_i^2 - m(2\bar{x} - m)$.
2. Задается доверительный уровень α и вычисляются $\beta_n = (1 - \alpha)/2$ и $\beta_n = (1 + \alpha)/2$.
3. Определяются $t_n = F_n^{-1}(\beta_n)$ и $t_n = F_n^{-1}(\beta_n)$, где F_n^{-1} — функция, обратная к функции распределения χ^2 с n степенями свободы.
4. Вычисляется доверительный интервал: $\left(\frac{n}{t_n} S_n^2, \frac{n}{t_n} S_n^2 \right)$.

Комментарии

1. Метод не устойчив при отклонении от нормальности.
2. Если математическое ожидание неизвестно, применяется метод построения доверительных интервалов из следующего раздела.
3. Знание точного значения математического ожидания в общем случае несущественно уменьшает длину доверительного интервала (по сравнению, например, с доверительным интервалом, построенным без использования точного значения математического ожидания). Поэтому, если есть сомнения в точном значении математического ожидания, следует использовать метод построения доверительных интервалов из раздела 10.3.4.

Практическая реализация

На рис. 10.6 показан рабочий лист Excel со всеми формулами, необходимыми для вычисления доверительного интервала. В качестве “подопытной” выборки используется выборка из предыдущих разделов. Напомним, что точное значение дисперсии равно 4. О точности доверительного интервала читатель может судить самостоятельно. Отметим, что для вычисления t_n и t_n здесь использована функция ХИ2ОБР (см. раздел 4.7.8).

10.3.4. Интервальные оценки для неизвестной дисперсии при неизвестном математическом ожидании

Математическое ожидание m и дисперсия σ^2 распределения генеральной совокупности неизвестны. Доверительный интервал для σ^2 строится следующим образом.

C10		=СУММКВ(Выборка)/С6-С2*(2*С8-С2)						
	A	B	C	D	E	F	G	H
1	Выборка	Известное мат. ожидание						
2	-1,60046	-1						
3	-3,55537	Доверительный уровень						
4	-0,51149	0,95						
5	1,552947	Объем выборки						
6	1,3967	50 =СЧЕТ(Выборка)						
7	2,466266	Среднее						
8	-5,36718	-1,214007 =СРЗНАЧ(Выборка)						
9	-1,46836	Выборочная дисперсия						
10	1,190045	5,351577 =СУММКВ(Выборка)/С6-С2*(2*С8-С2)						
11	-3,1734	t верхнее t нижнее						
12	-2,38041	71,42019 32,35738 ← =ХИ2ОБР((1+С4)/2,С6)						
13	-4,38086	← =ХИ2ОБР((1-С4)/2,С6)						
14	-4,69382							
15	-2,95526	Доверительный интервал						
16	-2,54701	3,746543 8,269483						
17	-5,23586	=С6*С10/С12 ← =С6*С10/Д12						
18	-2,13585							
19	-1,8081							
20	-0,73029							

Рис. 10.6. Построение доверительного интервала для дисперсии при известном математическом ожидании

1. Вычисляются точечные оценки $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ и $S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$.
2. Задается доверительный уровень α и вычисляются $\beta_n = (1 - \alpha)/2$ и $\beta_n = (1 + \alpha)/2$.
3. Определяются $t_n = F_{n-1}^{-1}(\beta_n)$ и $t_n = F_{n-1}^{-1}(\beta_n)$, где F_{n-1}^{-1} — функция, обратная к функции распределения χ^2 с $(n - 1)$ степенями свободы.
4. Вычисляется доверительный интервал: $\left(\frac{n-1}{t_n} S_n^2, \frac{n-1}{t_n} S_n^2 \right)$.

Комментарии

1. Метод не устойчив при отклонении от нормальности.
2. Если известно точное значение математического ожидания, можно использовать метод из предыдущего раздела.

Практическая реализация

На рис. 10.7 показан рабочий лист Excel со всеми формулами, необходимыми для вычисления доверительного интервала. Для примера используется выборка из предыдущих разделов. Напомним, что точное значение дисперсии равно 4. Для вычисления t_n и t_n здесь использована функция ХИ2ОБР (см. раздел 4.7.8).

C10		$\mu = \text{ХИ2ОБР}((1-C2)/2; C4-1)$					
	A	B	C	D	E	F	G
1	Выборка	Доверительный уровень					
2	-1.60046	0.95					
3	-3.55537	Объем выборки					
4	-0.51149	50 = СЧЕТ(Выборка)					
5	1.552947	Среднее					
6	1.3967	-1.21401 = СРЗНАЧ(Выборка)					
7	2.466266	Выборочная дисперсия					
8	-5.36718	5.305778 = ДИСПР(Выборка)					
9	-1.46836	t верхнее и нижнее					
10	1.190045	70.22236	31.55493 = ХИ2ОБР((1+C2)/2; C4-1)				
11	-3.1734	= ХИ2ОБР((1-C2)/2; C4-1)					
12	-2.38041	Доверительный интервал					
13	-4.38086						
14	-4.69382	3.702284	8.239065	= (C4-1)*C8/D10			
15	-2.95526	= (C4-1)*C8/C10					
16	-2.54701						
17	-5.23586						
18	-2.13585						
19	-1.8081						
20	-0.73029						

Рис. 10.7. Построение доверительного интервала для дисперсии при известном математическом ожидании

10.4. Оценка параметров логарифмически нормального распределения

Напомним, что если случайная величина X имеет логнормальное распределение, то ее логарифм $Y = \ln X$ распределен по нормальному закону с математическим ожиданием m и дисперсией σ^2 . Поэтому оценивание параметров m и σ^2 можно проводить точно так, как оценивание параметров m и σ^2 нормального распределения (см. раздел 10.3), если выборочные значения x_1, x_2, \dots, x_n заменить значениями $\ln x_1, \ln x_2, \dots, \ln x_n$. Например, построим доверительный интервал для параметра m , предполагая, что значение параметра σ^2 неизвестно.

Статистическая модель. Генеральная совокупность имеет логарифмически нормальное распределение с параметрами m и σ^2 (см. раздел 1.5.8).

1. Вычисляются точечные оценки $\bar{m} = \frac{1}{n} \sum_{i=1}^n \ln x_i$ и $S_n^2 = \frac{1}{n} \sum_{i=1}^n (\ln x_i - \bar{m})^2$.

2. Задается доверительный уровень α .

3. Из уравнения $\alpha = 2F_{n-1}(k) - 1$, где F_{n-1} — функция распределения Стюдента с $(n-1)$ степенями свободы, вычисляется значение k : $k = F_{n-1}^{-1}\left(\frac{1+\alpha}{2}\right)$,

F_{n-1}^{-1} — функция, обратная к функции распределения Стюдента.

4. Вычисляется доверительный интервал: $\left(\bar{m} - k \frac{S_n}{\sqrt{n-1}}, \bar{m} + k \frac{S_n}{\sqrt{n-1}}\right)$.

Комментарий. В сущности, здесь повторяется метод из раздела 10.3.2.

Практическая реализация

На рис. 10.8 показана выборка (столбец А), имеющая логарифмически нормальное распределение с параметрами $m = -1$ и $\sigma^2 = 4$. Выборка построена с помощью формулы массива

$$\{=\text{ЛОГНОРМОБР}(\text{СЛЧИС}();-1;2)\},$$

затем результаты вычисления по этой формуле преобразованы в значения (после копирования диапазона, содержащего вычисления, выполняется команда Правка → Специальная вставка → Значения). В Excel можно построить доверительный интервал, не вычисляя специально по выборочным значениям x_1, x_2, \dots, x_n значения $\ln x_1, \ln x_2, \dots, \ln x_n$. Для этого опять надо воспользоваться формулами массива. Для определения

выборочного значения параметра $\bar{m} = \frac{1}{n} \sum_{i=1}^n \ln x_i$ в ячейке В4 использована формула

массива $\{=\text{СРЗНАЧ}(\text{LN}(\text{Выборка}))\}$, для вычисления корня из величины

$S_n^2 = \frac{1}{n} \sum_{i=1}^n (\ln x_i - \bar{m})^2$ в ячейке В6 применена формула $\{=\text{СТАНДОТКЛОН}(\text{LN}(\text{Выборка}))\}$.

Подчеркнем необходимость использования именно формул массивов — их использование в качестве “обычных” формул приведет к неверным результатам. Остальные расчетные формулы в столбце D ничем не отличаются от аналогичных формул для построения доверительного интервала из раздела 10.3.2.

	B4		F4				
	A	B	C	D	E	F	G
1	Выборка	Объем выборки		Доверительный уровень			
2	0,87443	50		0,95			
3	0,151718	Среднее		Коэффициент k			
4	0,777133	-1,009073527		2,3123721	=СТЮДРАСПОБР((1-D2)/2;B2-1)		
5	0,151715	Ст. отклонение		Граница			
6	0,632726	1,605513382		0,5303635	=D4*В6/КОРЕНЬ(B2-1)		
7	0,054823	=СТАНДОТКЛОН(LN(Выборка))					
8	0,623124			Доверительный интервал			
9	0,104355			-1,539437	-0,4787101		
10	0,272998			=B4-D6	=B4+D6		
11	0,402482						
12	1,492359						
13	0,26699						
14	0,272248						
15	0,062516						
16	0,08594						

Рис. 10.8. Построение доверительного интервала для параметра m лог-нормального распределения

10.5. Оценка параметра показательного распределения

Напомним, что показательное (экспоненциальное) распределение определяется одним параметром λ (см. раздел 1.5.3), при этом для случайной величины X , подчиняющейся этому распределению, $MX = 1/\lambda$, $DX = 1/\lambda^2$. Для этого распре-

деления обычно оценивается не параметр λ , а обратная к нему величина $\theta = 1/\lambda$ (что естественно с учетом равенства $MX = \theta$). Построение доверительного интервала для параметра θ основано на том, что случайная величина $2 \sum_{i=1}^n x_i / \theta$, где x_i — выборочные значения, имеющие показательное распределение с параметром θ , не зависит от θ и имеет распределение χ^2 с $2n$ степенями свободы.

Статистическая модель. Генеральная совокупность имеет показательное распределение с параметром θ .

Доверительный интервал для θ строится следующим образом.

1. Вычисляется точечная оценка $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.
2. Задается доверительный уровень α и вычисляются $\beta_n = (1 - \alpha)/2$ и $\beta_n = (1 + \alpha)/2$.
3. Определяются $t_n = F_{2n}^{-1}(\beta_n)$ и $t_n = F_{2n}^{-1}(\beta_n)$, где F_{2n}^{-1} — функция, обратная к функции распределения χ^2 с $2n$ степенями свободы.
4. Вычисляется доверительный интервал: $\left(\frac{2n}{t_n} \bar{x}, \frac{2n}{t_n} \bar{x} \right)$.

Комментарии

1. Поскольку здесь математическое ожидание и дисперсия зависимы, для построения доверительного интервала нельзя использовать выборочную оценку дисперсии.
2. По той же причине интервал строится на основе выборочного среднего, а не выборочной дисперсии, которая с необходимостью будет использовать значение среднего.

Практическая реализация

На рис. 10.9 показана выборка (столбец А), имеющая показательное распределение с параметром $\lambda = 0,5$ (или $\theta = 2$). Выборка построена с помощью формулы массива `{=ГАММАОБР(СЛЧИС();1;2)}`. В Excel нет специальной функции, обратной к функции распределения показательного закона (есть функция ЭКСПРАСП, вычисляющая значения плотности вероятности и функции распределения), но, поскольку это распределение является частным случаем гамма-распределения при $\alpha = 1$, можно воспользоваться функцией ГАММАОБР (см. раздел 4.7.3), если положить в ней второй аргумент равным 1. Попутно отметим, что третий аргумент в этой функции задает параметр θ , а не λ . Все формулы, необходимые для построения доверительного интервала, показаны на рис. 10.9.

10.6. Оценка параметров гамма-распределения

Напомним, что гамма-распределение зависит от параметров α и λ ($\alpha > 0$, $\lambda > 0$), при этом $MX = \alpha/\lambda$, $DX = \alpha/\lambda^2$ (см. раздел 1.5.10). Если $\alpha = 1$, то гамма-распределение совпадает с показательным, оценки для которого рассмотрены

в предыдущем разделе. Так же, как и в случае показательного распределения, здесь вместо параметра λ оценивается обратный параметр $\theta = 1/\lambda$. Рассмотрим сначала вариант, когда известен параметр α .

C8		n = $\text{ХИ2ОБР}((1-C2)/2, 2*C4)$			
A	B	C	D	E	F
1	Выборка	Доверительный уровень			
2	3,716132	0,95			
3	0,986445	Объем выборки			
4	6,578394	50 = СЧЕТ(Выборка)			
5	0,339693	Среднее			
6	0,728646	2,126367 = СРЗНАЧ(Выборка)			
7	4,41567	t верхнее t нижнее			
8	1,42214	129,5613	74,22188 = $\text{ХИ2ОБР}((1+C2)/2, 2*C4)$		
9	5,520742	= $\text{ХИ2ОБР}((1-C2)/2, 2*C4)$			
10	3,860405				
11	11,77425	Доверительный интервал			
12	6,446062	1,641206	2,864879	= $2*C6*C4/D8$	
13	13,29812	= $2*C6*C4/C8$			
14	1,353292				
15	0,014298				
16	0,381732				
17	2,252559				
18	4,155054				

Рис. 10.9. Построение доверительного интервала для параметра показательного распределения

10.6.1. Оценка параметра λ при известном параметре α

Статистическая модель. Выборка x_1, x_2, \dots, x_n получена из генеральной совокупности, имеющей гамма-распределение с параметрами α и λ (см. раздел 1.5.10). Параметр α предполагается известным.

Доверительный интервал для $\theta = 1/\lambda$ строится следующим образом.

1. Вычисляется точечная оценка $\bar{\theta} = \frac{1}{n\alpha} \sum_{i=1}^n x_i$.
2. Задается доверительный уровень p и вычисляются $\beta_n = (1 - p)/2$ и $\beta_n = (1 + p)/2$.
3. Определяются $t_n = F^{-1}(\beta_n)$ и $t_n = F^{-1}(\beta_n)$, где F^{-1} — функция, обратная к функции гамма-распределения с параметрами $\alpha_1 = n\alpha$, $\lambda = 1$.
4. Вычисляется доверительный интервал: $\left(\frac{n\alpha\bar{\theta}}{t_n}, \frac{n\alpha\bar{\theta}}{t_n} \right)$.

Комментарии

1. Поскольку здесь математическое ожидание и дисперсия зависимы, для построения доверительного интервала нельзя использовать выборочную оценку дисперсии.

2. Доверительный интервал построен на основе того факта, что случайная величина $\frac{n\lambda\bar{\theta}}{\theta}$ также имеет гамма-распределение с параметрами $\alpha_1 = n\lambda$ и $\lambda = 1$, т.е. не зависит от неизвестного параметра θ .

Практическая реализация

На рис. 10.10 показана выборка (столбец А), имеющая гамма-распределение с параметрами $\alpha = 3$ и $\lambda = 0,5$ (или $\theta = 2$). Выборка построена с помощью формулы массива $\{=\text{ГAMMAOБP}(\text{СЛЧИС}();3;2)\}$. Все формулы, необходимые для построения доверительного интервала, показаны на рис. 10.10.

C7		=CPЗНАЧ(Выборка)/D1				
1	Выборка	Альфа=	3	Тетта=	2	
2	7,81406	Доверительный уровень				
3	10,0589	0,95				
4	4,124049	Объем выборки				
5	4,568642	50=СЧЕТ(Выборка)				
6	6,558207	Выборочное тетта				
7	10,65411	1,927708=CPЗНАЧ(Выборка)/D3				
8	4,000776	t верхнее t нижнее				
9	6,224809	174,9373 126,9581 =ГAMMAOБP((1-С\$3)/2,С\$5*С\$1,1)				
10	7,20629	=ГAMMAOБP((1+С\$3)/2,С\$5*С\$1,1)				
11	4,56916					
12	3,947125	Доверительный интервал				
13	3,83438	1,652914 2,277607 =С\$1*С\$5*С\$7/D9				
14	5,945972	=С\$1*С\$5*С\$7/C9				
15	4,245962					
16	2,053862					
17	2,667266					

Рис. 10.10. Построение доверительного интервала для параметра θ гамма-распределения

10.6.2. Оценка параметра α при известном параметре λ

Статистическая модель. Выборка x_1, x_2, \dots, x_n получена из генеральной совокупности, имеющей гамма-распределение с параметрами α и λ . Параметр λ предполагается известным.

Доверительный интервал для α строится следующим образом.

1. Вычисляется точечная оценка $\bar{\alpha} = \frac{\lambda}{n} \sum_{i=1}^n x_i$.
2. Задается доверительный уровень p и вычисляются $\beta_n = (1-p)/2$ и $\beta_n = (1+p)/2$.
3. Определяются $t_n = F^{-1}(\beta_n)$ и $t_n = F^{-1}(\beta_n)$, где F^{-1} — функция, обратная к функции гамма-распределения с параметрами $\alpha = n/\lambda$ и $\lambda = 1$.

4. Вычисляется доверительный интервал: $\left(\frac{n\bar{a}}{\lambda_{\alpha}}, \frac{n\bar{a}}{\lambda_{\alpha}} \right)$.

Комментарии

1. Поскольку здесь математическое ожидание и дисперсия зависимы, для построения доверительного интервала нельзя использовать выборочную оценку дисперсии.
2. Доверительный интервал построен на основе того факта, что случайная величина $\frac{\bar{a}}{aX}$ также имеет гамма-распределение с параметрами $a_x = ga/A$ и $A = 1$, т.е. не зависит от неизвестного параметра a .

Практическая реализация

На рис. 10.11 показана выборка (столбец А), имеющая гамма-распределение с параметрами $a = 3$ и $X = 0,5$ (или $9 = 2$). Выборка построена с помощью формулы массива $\{=\text{ГАММАОБР}(\text{СЛЧИС}();3;2)\}$. Все формулы, необходимые для построения доверительного интервала, показаны на рис. 10.11.

С7		=D1*СРЗНАЧ(Выборка)	
А	В		
1. Выборка	Лямбда	0,5	Альфа= 3
2. 5,222141	Доверительный уровень		
3. 14,37649	0,95		
4. 8,445031	Объем выборки		
5. 5,653292	50	=СЧЕТ(Выборка)	
6. 9,404357	Выборочное альфа		
7. 6,350556	2,870443	=D1*СРЗНАЧ(Выборка)	
8. 3,591874	t верхнее t нижнее		
9. 2,19432	120,529 81,36398	=ГАММАОБР((1-С\$3)/2,С\$5/\$D\$1;1)	
10. 6,346354		=ГАММАОБР((1+С\$3)/2,С\$5/\$D\$1;1)	
11. 11,34846			
12. 4,20127	Доверительный интервал		
13. 9,707628	2,381537 3,527904	=С\$5*С\$7/(\$D\$1*С9)	
14. 6,790415		=С\$5*С\$7/(\$D\$1*С9)	
15. 5,644929			
16. 2,489878			
	10,21881		

Рис. 10.11. Построение доверительного интервала для параметра α гамма-распределения

10.6.3. Совместная оценка параметров α и λ

Если неизвестны оба параметра (α и λ), то простых методов получения их оценок не существует. Поскольку для данного распределения $MX = \alpha/\lambda$ и $DX = \alpha/\lambda^2$, на основе значений выборочного среднего \bar{x} и выборочной дисперсии s_x^2 можно получить оценки этих параметров: $\bar{\lambda} = \frac{\bar{x}}{s_x^2}$ и $\bar{\alpha} = \bar{\lambda} \bar{x}$. Однако “теория” советует

использовать соотношения $MX = \alpha/\lambda$ и $MX^2 = \alpha(1 + \alpha)/\lambda^2$. На основе значений выборочного среднего \bar{x} и выборочного второго момента $\bar{m}_2 = \frac{1}{n} \sum_{i=1}^n x_i^2$ находим

$$\bar{\alpha} = \frac{\bar{x}^2}{\bar{m}_2 - \bar{x}^2}, \quad \bar{\lambda} = \frac{\bar{x}}{\bar{m}_2 - \bar{x}^2}.$$

Построить совместную *доверительную область* для этих оценок весьма сложно [17, раздел 6.4]. Можно, конечно, значение $\bar{\lambda}$ принять за точное значение параметра λ и построить доверительный интервал для α , как показано выше. Однако точность такого интервала оценить невозможно.

10.7. Оценка параметров равномерного распределения

В общем случае равномерное распределение задается границами конечного интервала $[a, b]$, на котором сосредоточено это распределение (см. раздел 1.5.1). Если границы не известны, то возникает следующая задача: по выборочным значениям, имеющим равномерное распределение, оценить значения границ.

Сначала рассмотрим случай, когда неизвестна только одна граница.

10.7.1. Оценка границы равномерного распределения

В этом случае с помощью линейной замены исходное распределение можно привести к распределению, сосредоточенному на интервале $[0, \theta]$, где θ — неизвестный параметр.

Статистическая модель. Выборка x_1, x_2, \dots, x_n получена из генеральной совокупности, имеющей равномерное распределение на интервале $[0, \theta]$. Параметр θ предполагается неизвестным.

Доверительный интервал для θ строится следующим образом.

1. Вычисляется максимальное выборочное значение $x_n^* = \max_{1 \leq i \leq n} x_i$.

2. Несмещенной и эффективной оценкой для параметра θ будет статистика

$$\bar{\theta} = \frac{n+1}{n} x_n^*. \quad (\text{Дисперсия этой статистики равна } D\bar{\theta} = \theta^2/n(n+2).)$$

3. Задается доверительный уровень α и вычисляется $\sqrt[n]{1-\alpha}$.

4. Вычисляется доверительный интервал: $\left(x_n^*, \frac{x_n^*}{\sqrt[n]{1-\alpha}} \right)$.

Комментарий. Доверительный интервал построен на основе того факта, что случайная величина $\bar{\theta}/\theta$ имеет распределение, не зависящее от параметра θ . Ее функция распределения задается формулой [8]

¹ Эти оценки соответствуют оценкам метода моментов. Их используют как начальное приближение для итерационного процесса нахождения оценок по методу максимального правдоподобия [17, раздел 6.4].

$$F(u) = \begin{cases} 0, & \text{если } u \leq 0, \\ \left(\frac{un}{n+1}\right)^2, & \text{если } u \in \left[0, 1 + \frac{1}{n}\right], \\ 1, & \text{если } u > 1 + \frac{1}{n}. \end{cases}$$

Несложно найти в явном виде корень уравнения $P(\bar{\theta}/\theta > u) = 1 - F(u) = \alpha$:

$u = \frac{n+1}{n} \sqrt{1-\alpha}$. Отсюда получаем границы доверительного интервала.

Практическая реализация построения в Excel доверительного интервала не представляет трудностей. Если выборочные значения записаны в диапазоне ячеек с именем Выборка, а значения доверительного уровня α — в ячейке с именем Альфа, то оценка $\bar{\theta}$ вычисляется по формуле

$$= \text{МАКС}(\text{Выборка}) * (\text{СЧЁТ}(\text{Выборка}) + 1) / \text{СЧЁТ}(\text{Выборка}),$$

нижняя граница доверительного интервала:

$$= \text{МАКС}(\text{Выборка}),$$

верхняя граница доверительного интервала:

$$= \text{МАКС}(\text{Выборка}) / \text{СТЕПЕНЬ}(1 - \text{Альфа}; 1 / \text{СЧЁТ}(\text{Выборка})).$$

10.7.2. Оценка обеих границ равномерного распределения

Статистическая модель. Выборка x_1, x_2, \dots, x_n получена из генеральной совокупности, имеющей равномерное распределение на интервале $[a, b]$ с неизвестными параметрами a и b . Предполагается, что $0 < a < b$.

Несмещенными и эффективными оценками для параметров a и b будут соответственно оценки

$$\bar{a} = \frac{nx_1^*}{n-1} - \frac{x_n^*}{n-1}, \quad \bar{b} = \frac{nx_n^*}{n-1} - \frac{x_1^*}{n-1},$$

где $x_1^* = \min_{1 \leq i \leq n} x_i$, $x_n^* = \max_{1 \leq i \leq n} x_i$. Совместную доверительную область для этих оценок построить сложно.

Приведем еще несмещенную и эффективную оценку для размаха $R = b - a$:

$$\bar{R} = \frac{n+1}{n-1} (x_n^* - x_1^*).$$

Практическая реализация в Excel приведенных формул несложна и очевидна.

10.8. Оценки параметра распределения Бернулли

Напомним, что распределение Бернулли обычно рассматривается как модель случайного эксперимента, в результате которого с вероятностью p может произойти исход "1" и с вероятностью $(1-p)$ — исход "0" (см. раздел 1.4.2). Как правило, целью статистического анализа выборочных значений является определение значения биномиальной вероятности p . Общие теоретические положения, на основе которых строятся описываемые ниже доверительные интервалы, приведены в разделе 2.3.7.

Построение доверительного интервала для вероятности p несколько отличается для случаев, когда выборка содержит наблюдения за одним экспериментом и когда выборка содержит результаты нескольких независимых экспериментов. Рассмотрим эти случаи отдельно.

10.8.1. Оценивание вероятности p по одному эксперименту

Статистическая модель. Выборка x_1, x_2, \dots, x_n является результатом наблюдения за одним экспериментом, состоящим из n одинаковых испытаний, в каждом из которых с вероятностью p может произойти исход "1" и с вероятностью $(1-p)$ — исход "0". Здесь x_i равно 1, если в i -м испытании произошел исход "1", и 0 в противном случае.

Несмещенной и эффективной оценкой для вероятности p будет статистика $\hat{p} = r/n$, где r — количество исходов "1". Случайная величина r имеет биномиальное распределение с параметрами n и p (см. раздел 1.4.3). Распределение статистики \hat{p} асимптотически нормально с параметрами $m = p$ и $\sigma^2 = p(1-p)/n$.

Доверительные интервалы для неизвестного значения вероятности p строятся или на основе биномиального распределения, которое имеет случайная величина r , или на основе асимптотической нормальности распределения статистики \hat{p} .

Использование биномиального распределения

Доверительный интервал для значения вероятности p строится следующим образом.

1. Вычисляется точечная оценка $\hat{p} = r/n$, точнее, для дальнейших вычислений необходима величина r — количество исходов "1".
2. Задается доверительный уровень α и вычисляются $\beta_\alpha = (1 - \alpha)/2$ и $\beta_\alpha = (1 + \alpha)/2$.
3. Определяются $t_n = F_{k1, k2}^{-1}(\beta_\alpha)$ и $t_n = F_{k3, k4}^{-1}(\beta_\alpha)$, где $F_{m1, m2}^{-1}$ — функция, обратная к функции F -распределения с параметрами $m1$ и $m2$ (см. раздел 1.5.9), $k1 = r$, $k2 = n - r + 1$, $k3 = r + 1$, $k4 = n - r$.
4. Доверительным интервалом будет интервал (t_n, t_n) .

Комментарий. Здесь использованы известные соотношения между биномиальным распределением и F -распределением. Пусть X — случайная величина, имеющая биномиальное распределение с параметрами n и p . Тогда $P(X \leq k) = F_{n-k, k+1}(1-p)$, где $F_{n-k, k+1}$ — функция F -распределения с соответствующими параметрами.

Практическая реализация

На рис. 10.12 в столбце А показана выборка, содержащая 100 наблюдений за экспериментом, где с вероятностью 0,4 происходит событие "1". Эта выборка получена с помощью средства пакета анализа Генерация случайных чисел. Все формулы Excel, необходимые для построения доверительного интервала, также показаны на рис. 10.12. Можно обойтись без промежуточных вычислений, применив для вычисления нижней границы доверительного интервала формулу

$$= \text{ФРАСПОБР}((1-C2)/2; \text{СУММ(Выборка)}; \text{СЧЕТ(Выборка)} - \text{СУММ(Выборка)} + 1),$$

а для верхней — формулу

=ФРАСПОБР((1+C2)/2;СУММ(Выборка)+1;СЧЕТ(Выборка)-СУММ(Выборка)).

Эти формулы используют только выборочные значения (диапазон Выборка) и значение доверительного уровня (ячейка C2).

C10			= (1-C2)/2
Выборка	Доверительный уровень	k1=	46 = C6
1	0,95	k2=	55 = C4-C6+1
1	Объем выборки	k3=	47 = C6+1
0	100 = СЧЕТ(Выборка)	k4=	54 = C4-C6
0	r		
0	46 = СУММ(Выборка)		
0	Точечная оценка p		
1	0,46 = СУММ(Выборка)/СЧЕТ(Выборка)		
0	Бета нижнее Бета верхнее		
0	0,025 0,975 = (1+C2)/2		
1	Доверительный интервал		
1	t нижнее t верхнее		
1	0,359843254 0,562588215	= ФРАСПОБР(D8;F3;F4)	
1	= ФРАСПОБР(C8;F1;F2)		
1			
1			

Рис. 10.12. Построение доверительного интервала для параметра p

Асимптотические оценки

При достаточно большом n ($n \geq 30$) приближенный доверительный интервал для значения вероятности p строится таким образом.

1. Вычисляется точечная оценка $\hat{p} = r/n$.
2. Задается доверительный уровень α .
3. Из уравнения $\alpha = 2\Phi(k) - 1$, где Φ — функция распределения стандартного нормального закона, определяется значение k : $k = \Phi^{-1}\left(\frac{1+\alpha}{2}\right)$, Φ^{-1} — функция, обратная к функции распределения стандартного нормального закона.
4. Строится доверительный интервал вида

$$\left(\hat{p} - k \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + k \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \right).$$

Комментарий. При построении доверительного интервала биномиальное распределение аппроксимируется нормальным, неизвестное значение дисперсии $D\hat{p} = p(1-p)/n$ заменяется величиной $\hat{p}(1-\hat{p})/n$.

Практическая реализация

На рис. 10.13 показаны формулы Excel, позволяющие построить асимптотический доверительный интервал. Здесь использована та же выборка, что и в предыдущем примере, и оставлены вычисления точного доверительного интервала. Как видно на рис. 10.13, асимптотический доверительный интервал (ячейки H13 и I13), на первый взгляд, кажется более точным, чем интервал, построенный по точным формулам. Однако необходимо помнить, что этот интервал приближенный и он может быть как меньше, так и больше точного интервала.

H7		=СУММ(Выборка)/СЧЕТ(Выборка)			
1	Выборка	Точные оценки		Асимптотические оценки	
2	1	Доверительный уровень	k1= 46	Доверительный уровень	
3	1	0,95	k2= 55	0,95	
4	0	Объем выборки	k3= 47	Объем выборки	
5	0	100	k4= 54	100	
6	0	г		Точечная оценка p	
7	0	46		0,46	
8	1	Точечная оценка p		Коэффициент k	
9	0	0,46		1,959963 =НОРМСТОБР((1+H3)/2)	
10	0	Бета нижнее Бета верхнее		Граница	
11	1	0,025 0,975		0,097684 =H9*КОРЕНЬ(H7*(1-H7)/H5)	
12	1	Доверительный интервал		Доверительный интервал	
13	1	t нижнее t верхнее		0,362316 0,557684	
14	1	0,359843254 0,562588215		=H7-H11 =H7+H11	
15	1				
16	1				

Рис. 10.13. Асимптотический доверительный интервал для параметра p

10.8.2. Оценивание вероятности p по нескольким экспериментам

Статистическая модель. Выборка x_1, x_2, \dots, x_n состоит из результатов n экспериментов, в каждом из которых проводилось N испытаний, в каждом из которых с вероятностью p может произойти исход "1" и с вероятностью $(1 - p)$ — исход "0". Здесь x_i равно числу исходов "1" в i -м эксперименте.

Несмещенной и эффективной оценкой для вероятности p будет статистика

$$\hat{p} = \frac{1}{nN} \sum_{i=1}^n x_i. \text{ Дисперсия статистики } \hat{p}: D\hat{p} = p(1-p)/nN. \text{ Распределение статисти}$$

стики \hat{p} асимптотически нормально с параметрами $m = p$ и $\sigma^2 = p(1-p)/nN$.

Поскольку значение величины nN , как правило, больше 30, наиболее простой доверительный интервал для неизвестного значения вероятности p строится на основе асимптотической нормальности распределения статистики \hat{p} .

1. Вычисляется точечная оценка $\hat{p} = \frac{1}{nN} \sum_{i=1}^n x_i$.

2. Задается доверительный уровень α .

3. Из уравнения $\alpha = 2\Phi(k) - 1$, где Φ — функция распределения стандартного нормального закона, находится значение k : $k = \Phi^{-1}\left(\frac{1+\alpha}{2}\right)$, Φ^{-1} — функция, обратная к функции распределения стандартного нормального закона.

4. Вычисляется доверительный интервал $\left(\hat{p} - k\sqrt{\frac{\hat{p}(1-\hat{p})}{nN}}, \hat{p} + k\sqrt{\frac{\hat{p}(1-\hat{p})}{nN}} \right)$.

Комментарии

1. При необходимости можно построить точный доверительный интервал, аналогичный точному интервалу из предыдущего раздела. Но поскольку значение величины nN , как правило, велико, на практике обычно используют асимптотические интервалы как наиболее простые в вычислительном отношении и вместе с тем достаточно надежные.
2. Здесь при построении доверительного интервала используется аппроксимация биномиального распределения нормальным, а неизвестное значение дисперсии $D\hat{p}$ заменяется величиной $\hat{p}(1-\hat{p})/nN$.
3. В случае, когда в экспериментах проводится разное количество испытаний N_1, N_2, \dots, N_n , все вышеприведенные формулы сохраняют свою силу, если в них величину nN заменить суммой $N_1 + N_2 + \dots + N_n$.

Практическая реализация этого метода построения доверительного интервала с небольшими очевидными изменениями повторяет реализацию метода построения асимптотического доверительного интервала из предыдущего раздела.

10.8.3. Применение преобразования арксинуса

В разделе 2.3.7 описаны преобразования оценки \hat{p} вида $z = \arcsin\sqrt{\hat{p}}$ и $y = 2\sqrt{n} \arcsin\sqrt{\hat{p}}$, при этом распределения величин z и y ближе к нормальному, чем распределение оценки \hat{p} . Напомним, что математическое ожидание случайной величины z приближенно равно $\arcsin\sqrt{p}$, а дисперсия приближенно равна $1/4n$. Для величины y математическое ожидание приближенно равно $2\sqrt{n} \arcsin\sqrt{p}$, дисперсия приближенно равна 1. Эти преобразования можно использовать для построения доверительного интервала для вероятности p .

Предварительно отметим, что если дисперсию величины z принять равной в точности $1/4n$, а величины y — точно равной 1, то, как нетрудно проверить, доверительные интервалы, построенные на основании этих преобразований, будут совпадать. Поэтому не имеет значения, какое преобразование использовать. Покажем построение доверительного интервала с помощью преобразования арксинуса $z = \arcsin\sqrt{\hat{p}}$.

При построении доверительного интервала выполняются следующие действия.

1. Вычисляются точечная оценка \hat{p} и ее преобразование $\bar{z} = \arcsin\sqrt{\hat{p}}$.
2. Задается доверительный уровень α .

3. Из уравнения $\alpha = 2\Phi(k) - 1$, где Φ — функция распределения стандартного нормального закона, вычисляется значение k : $k = \Phi^{-1}\left(\frac{1+\alpha}{2}\right)$, Φ^{-1} — функция, обратная к функции распределения стандартного нормального закона.

4. Вычисляется доверительный интервал: $\left(\sin^2\left(\bar{z} - \frac{k}{2\sqrt{n}}\right), \sin^2\left(\bar{z} + \frac{k}{2\sqrt{n}}\right)\right)$.

Комментарий. Необходимо помнить, что, во-первых, хотя распределение величины z ближе к нормальному, чем распределение \hat{p} , оно все-таки не совпадает с ним. Во-вторых, дисперсия этой величины только приблизительно равна $1/4n$. Поэтому данный метод является приближенным.

Практическая реализация

На рис. 10.14 показаны формулы Excel, позволяющие построить доверительный интервал на основании преобразования $z = \arcsin \sqrt{\hat{p}}$. Здесь использована та же выборка, что и в разделе 10.8.1. Как видно на рис. 10.14, построенный доверительный интервал (ячейки F6 и G6) близок асимптотическому доверительному интервалу, построенному в разделе 10.8.1.

	A	B	C	D	E	F	G	H
1	Выборка	Доверительный уровень	Преобразование					
2	1	0,95				z	0,745355	=ASIN(KOPEHЬ(D5))
3	1	Объем выборки	100			Интервал для z		
4	0	г	46			0,647357	0,843354	=\$G\$2+\$D\$6/(2*KOPEHЬ(\$D\$3))
5	0	Точечная оценка p	0,46			Интервал для p		
6	0	Коэффициент k	1,96			0,363706	0,557826	=SIN(G4)*SIN(G4)
7	0	=НОРМСТОБР((1+C2)/2)						=SIN(F4)*SIN(F4)
8	1							
9	0							=\$G\$2-\$D\$6/(2*KOPEHЬ(\$D\$3))
10	0							
11	1							
12	1							

Рис. 10.14. Доверительный интервал для параметра p на основе преобразования арксинуса

10.9. Оценка параметра распределения Пуассона

Статистическая модель. Генеральная совокупность имеет распределение Пуассона с параметром λ (см. раздел 1.4.4).

Выборочное среднее $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ будет несмещенной и эффективной оценкой для неизвестного параметра λ . Дисперсия этой оценки равна $D\bar{x} = \lambda/n$. Случайная величина $\sum_{i=1}^n x_i$ имеет распределение Пуассона с параметром $n\lambda$, а случайная

величина $\sqrt{\frac{n}{\lambda}}(\bar{x} - \lambda)$ асимптотически нормальна с параметрами $(0, 1)$.

Доверительные интервалы для параметра λ строятся или на основе распределения Пуассона, которое имеет случайная величина $\sum_{i=1}^n x_i$, или на основе асимптотической нормальности распределения случайной величины $\sqrt{\frac{n}{\lambda}}(\bar{x} - \lambda)$.

Использование распределения Пуассона

Доверительный интервал для параметра λ строится следующим образом.

1. Вычисляется точечная оценка $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.
2. Задается доверительный уровень α и вычисляются $\beta_{\alpha} = (1 - \alpha)/2$ и $\beta_{\alpha} = (1 + \alpha)/2$.
3. Определяются $t_{\alpha} = F_k^{-1}(\beta_{\alpha})$ и $t_{\alpha} = F_k^{-1}(\beta_{\alpha})$, где F_k^{-1} — функция, обратная к функции χ^2 -распределения с $k = 2(n\bar{x} + 1)$ степенью свободы.
4. Доверительным интервалом будет интервал $\left(\frac{t_{\alpha}}{2n}, \frac{t_{\alpha}}{2n}\right)$.

Комментарий. Здесь использованы известные соотношения между распределением Пуассона и распределением χ^2 , приведенные в разделе 1.4.4.

Практическая реализация

Реализация этого метода построения доверительного интервала с соответствующими формулами показана на рис. 10.15. В столбце А содержатся 100 выборочных значений, имеющих распределение Пуассона с параметром $\lambda = 2$. Выборка создана с помощью средства Генерация случайных чисел; диапазону ячеек, содержащему выборочные значения, присвоено имя Выборка. Отметим, что для вычисления t_{α} и t_{α} здесь использована функция ХИ2ОБР (см. раздел 4.7.8).

Асимптотические оценки

При достаточно большом n ($n \geq 30$) приближенный доверительный интервал для значения λ строится таким образом.

1. Вычисляется точечная оценка $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.
2. Задается доверительный уровень α .
3. Из уравнения $\alpha = 2\Phi(k) - 1$, где Φ — функция распределения стандартного нормального закона, определяется значение k : $k = \Phi^{-1}\left(\frac{1+\alpha}{2}\right)$, Φ^{-1} — функция, обратная к функции распределения стандартного нормального закона.
4. Строится доверительный интервал: $\left(\bar{x} - k\sqrt{\frac{\bar{x}}{n}}, \bar{x} + k\sqrt{\frac{\bar{x}}{n}}\right)$.

C10		=ХИ2ОБР((1+C2/2,C8)
Выборка	Доверительный уровень	
2	0,95	
0	Объем выборки	
0	100 = СЧЕТ(Выборка)	
3	Среднее	
3	2,06 = СРЗНАЧ(Выборка)	
3	Степень свободы	
4	414 = 2*(C4*C6+1)	
1	t нижнее t верхнее	
2	359,5197 472,2676 = ХИ2ОБР((1-C2)/2,C8)	
2	=ХИ2ОБР((1+C2)/2,C8)	
0		
2	Доверительный интервал	
2	1,797598 2,361338 = D10/(2*C4)	
3	=C10/(2*C4)	
0		
2		

Рис. 10.15. Построение точного доверительного интервала для параметра λ распределения Пуассона

Комментарии

1. При построении доверительного интервала используется аппроксимация распределения Пуассона нормальным, при этом неизвестное значение дисперсии $Dx = X/n$ заменяется величиной x/n .
2. Можно строить доверительный интервал вида

$$\left(\bar{x} + \frac{k^2}{2n} - \frac{k}{2n} \sqrt{k^2 + 4n\bar{x}}, \bar{x} + \frac{k^2}{2n} + \frac{k}{2n} \sqrt{k^2 + 4n\bar{x}} \right),$$

где используется только аппроксимация распределения Пуассона нормальным.

3. Этот метод построения доверительного интервала является приближенным; по возможности следует использовать точный метод.

Практическая реализация

Все формулы Excel, необходимые для построения асимптотического доверительного интервала, показаны на рис. 10.16 в столбцах G и H. Для примера используется та же выборка, что и в предыдущем примере.

10.10. Оценки параметра геометрического распределения

Статистическая модель. Генеральная совокупность имеет геометрическое распределение с параметром p ($0 < p < 1$) (см. раздел 1.4.5).

При $n > 10$ доверительный интервал Для значения p строится таким образом.

1. Вычисляется точечная оценка $\bar{p} = \frac{1}{1 + \bar{x}}$, где $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$

G7		=CP3H4C(Выборка)					
	A	B	C	D	E	F	G
1	Выборка	Точные оценки			Асимптотические оценки		
2	2	Доверительный уровень			Доверительный уровень		
3	0	0,95			0,95		
4	0	Объем выборки			Объем выборки		
5	3	100			100		
6	3	Среднее			Среднее		
7	3	2,06			2,06		
8	4	Степень свободы			Коэффициент k		
9	1	414			1,959963 =НОРМСТОБР((1+G3)/2)		
10	2	t нижнее t верхнее			Граница		
11	2	359,5197 472,2676			0,281308 =G9*КОРЕНЬ(G7/G5)		
12	0				Доверительный интервал		
13	2	Доверительный интервал			1,778692 2,341308		
14	2	1,797599 2,361338			=G7-G11 =G7+G11		
15	3						
16	0						
17	2						

Рис. 10.16. Построение асимптотического доверительного интервала для параметра распределения Пуассона

2. Задается доверительный уровень α .
3. Из уравнения $\alpha = 2\Phi(k) - 1$, где Φ — функция распределения стандартного нормального закона, определяется значение k : $k = \Phi^{-1}\left(\frac{1+\alpha}{2}\right)$, Φ^{-1} — функция, обратная к функции распределения стандартного нормального закона.

4. Строится доверительный интервал: $\left(\frac{1}{1+\bar{x}} \left(1 - \frac{k}{\sqrt{n}} \sqrt{\frac{\bar{x}}{1+\bar{x}}} \right), \frac{1}{1+\bar{x}} \left(1 + \frac{k}{\sqrt{n}} \sqrt{\frac{\bar{x}}{1+\bar{x}}} \right) \right)$.

Комментарии

1. При построении доверительного интервала используется тот факт, что случайная величина $T = \frac{\sqrt{n(p\bar{x} - 1 + p)}}{\sqrt{1-p}}$ имеет асимптотически стандартное нормальное распределение. Если приравнять T к квантилям t_n и t_p соответственно порядка $\alpha/2$ и порядка $1 - \alpha/2$ стандартного нормального распределения, то получим два уравнения относительно p . Корни этих уравнений составляют границы доверительного интервала. Отметим, что приведенные выше границы получены при замене в формуле T выражения $\sqrt{1-p}$ выражением $\sqrt{\bar{x}/(1+\bar{x})}$. Без последней замены можно получить более точные границы, однако формулы для них становятся весьма громоздкими.
2. Этот метод построения доверительного интервала является приближенным, но уже при $n > 10$ дает удовлетворительную точность. Для малых выборок можно применить точный метод, использующий для вычисления границ доверительного интервала отрицательное биномиальное распределение [14].

Практическая реализация в Excel этого метода подобна построению асимптотического доверительного интервала для параметра λ распределения Пуассона из предыдущего раздела и не должна вызывать затруднений.

10.11. Доверительные интервалы для квантилей

В этом разделе рассмотрим доверительные интервалы для квантилей. Эти интервалы характерны тем, что не зависят от выборочного распределения. Напомним, что квантилью порядка p случайной величины X называется такое число ξ_p , что $P(X < \xi_p) = F(\xi_p) = p$, где F — функция распределения случайной величины X (см. раздел 1.2.3).

Статистическая модель. Выборка x_1, x_2, \dots, x_n получена из генеральной совокупности, имеющей непрерывное распределение.

Состоятельной оценкой для квантили ξ_p является порядковая статистика (член вариационного ряда) $x_{(k(p))}$ с рангом $k(p)$; $k(p) = np$, если np — целое число и $k(p) = [np] + 1$ в противном случае ($[np]$ — целая часть числа np).

Доверительный интервал для значения ξ_p строится следующим образом.

1. Для всех выборочных значений x_1, x_2, \dots, x_n вычисляются ранги r_1, r_2, \dots, r_n .
2. Вычисляется $k(p)$: $k(p) = np$, если np — целое число; $k(p) = [np] + 1$ в противном случае.
3. Определяется выборочное значение $x_{(k(p))}$ с рангом, равным $k(p)$. Это значение принимается за точечную оценку квантили ξ_p .
4. В качестве доверительного интервала берется интервал $(x_{(s)}, x_{(t)})$, границы которого составляют порядковые статистики $x_{(s)}$ и $x_{(t)}$ и который содержит значение $x_{(k(p))}$. Вычисляется доверительный уровень этого интервала по формуле $\beta = F(t) - F(s)$, где F — функция биномиального распределения с параметрами n и p . Порядковые статистики $x_{(s)}$ и $x_{(t)}$ выбираются таким образом, чтобы вероятность β была не меньше заданного доверительного уровня α .

Комментарии

1. Метод построен на основе свойств порядковых статистик (см. например, [17]).
2. Обычно интервал $(x_{(s)}, x_{(t)})$ берут симметричным по рангам s и t относительно $x_{(k(p))}$, т.е. когда $t - k(p) = k(p) - s$. Однако такой интервал не всегда имеет минимальную длину.
3. В случае $p = 0,5$ строится доверительный интервал для медианы.

Практическая реализация

Все формулы Excel, используемые при построении доверительного интервала для квантили ξ_p , показаны на рис. 10.17. Для примера используется выборка, состоящая из равномерно распределенных на интервале $[0, 10]$ случайных чисел (в этом случае истинное значение квантили ξ_p равно $10p$).

В данном случае для того, чтобы в дальнейшем можно было применить функцию ВПР, выборка записана в столбце В, а в столбце А вычислены ранги выборочных значений по формуле массива $\{=\text{РАНГ}(\text{Выборка};\text{Выборка};1)\}$. (Функция РАНГ описана в разделе 4.2.5; диапазон ячеек, содержащий выбо-

точные значения, назван Выборка.) В ячейке E3 вычисляется ранг ~~ранг~~ значения, которое принимается за оценку квантили, а в ячейке ~~ранг~~ вычисляется само это значение.

A2		{=РАНГ(Выборка;Выборка;1)}			
Ранги	Выборка	Порядок квантили	0,3		
27	5,074985	Количество	50		
20	3,543108	Ранг квантили	15=ЦЕЛОЕ(E1*E2)		
44	9,976215	Оценка квантили	3,102008581=ВПР(E3;A2:B51;2;0)		
38	7,519681	Интервалы			
13	2,898717	Номер	Левая граница	Правая граница	Вероятность
3	0,4171	1	3,027948165	3,21633996	0,237047
41	8,456682	2	2,898717425	3,35811432	0,4543088
28	5,185787	3	2,1401062	3,438827874	0,6365743
6	1,724152	4	2,1025591	3,535393657	0,7781613
26	5,005228	5	2,050231438	3,543107928	0,8733855
40	8,440815	6	2,023959372	3,643174955	0,9346813
33	7,384128	7	1,975008963	4,117563232	0,9694705
49	9,643886	8	1,734499439	4,188712678	0,9871436
50	9,901813	9	1,724151508	4,237482621	0,9951387
37	7,918145	10	1,346195778	4,841821597	0,998344
22	4,117583	11	0,701774179	5,005226265	0,9994888
4	0,701774	12	0,417100439	5,074985219	0,999852
24	4,237483	=ВПР(E\$3+C18;\$A\$2:\$B\$51;2;0)			
39	8,432058	=ВПР(E\$3-C18;\$A\$2:\$B\$51;2;0)			

Рис. 10.17. Построение доверительного интервала для квантили

В диапазоне C6:F18 построены интервалы для квантили и подсчитана вероятность, с которой они содержат неизвестное значение квантили. В ячейке этого диапазона записаны такие формулы.

В ячейке D7: =ВПР(\$E\$3-C7;\$A\$2:\$B\$51;2;0).

В ячейке E7: =ВПР(\$E\$3+C7;\$A\$2:\$B\$51;2;0).

В ячейке F7: =БИНОМРАСП(\$E\$3+C7;\$E\$2;\$E\$1;1)-БИНОМРАСП(\$E\$3-C7;\$E\$2;\$E\$1;1).

Эти формулы затем скопированы вниз до конца диапазона D7:F18.

Интервалы симметричны относительно оценки квантили (симметричность интервалов определена в комментариях). Столбец Номер используется только для удобства вычисления рангов z и t ; путем усложнения формул от него можно отказаться.

При заданном доверительном уровне выбирается тот интервал, для которого вероятность не меньше доверительного уровня. Например, если доверительный уровень задан как 0,95, то в качестве искомого интервала следует взять интервал под номером 7 (см. рис. 10.17).

Гипотезы

а) Равенство

б) Неравенство

в) Неравенство

$H_0: \mu = m_0$

$H_0: \mu \leq m_0$

$H_0: \mu \geq m_0$

$H_1: \mu \neq m_0$

$H_1: \mu > m_0$

$H_1: \mu < m_0$

Здесь m_0 — заданное число. Задан уровень значимости α .

Вычисления. Вычисляется критериальная статистика $T = \frac{\sqrt{n}(\bar{x} - m_0)}{S_n}$, где

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \text{ и } S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Построение критической области. При условии истинности гипотезы H_0 статистика T имеет распределение Стьюдента с $(n - 1)$ степенью свободы.

Случай а). Вычисляется критическое значение как квантиль t порядка $1 - \alpha/2$ распределения Стьюдента с $(n - 1)$ степенью свободы. Гипотеза H_0 принимается, если выполняется неравенство $|T| \leq t$, иначе гипотеза H_0 отклоняется.

Случай б). Вычисляется критическое значение как квантиль t_α порядка $1 - \alpha$ распределения Стьюдента с $(n - 1)$ степенью свободы. Гипотеза H_0 принимается, если $T \leq t_\alpha$.

Случай в). Вычисляется критическое значение как квантиль t_α порядка α распределения Стьюдента с $(n - 1)$ степенью свободы. Гипотеза H_0 принимается, если $t_\alpha \leq T$.

Комментарии

1. Иногда известна дисперсия генеральной совокупности σ^2 . Тогда вместо распределения Стьюдента используют стандартное нормальное распределение, а в формуле вычисления статистики T заменяют S_n на σ . В этом случае для построения критерия можно использовать функцию Excel ZTEST.

2. Критерий устойчив при умеренных отклонениях распределения выборки от нормального.

Практическая реализация

На рис. 11.1 показан рабочий лист Excel с формулами, необходимыми для реализации критерия. В качестве тестовой выборки взята выборка объемом 50 значений, имеющая нормальное распределение с математическим ожиданием -1 и дисперсией $\sigma^2 = 4$, полученная с помощью средства Генерация случайных чисел. Значение m_0 задается в ячейке C1, а уровень значимости критерия — в ячейке C2. Объем выборки, выборочные среднее и стандартное отклонение вычисляются с помощью функций СЧЁТ, СРЗНАЧ и СТАНДОТКЛОН.

В столбцах E:G выполняются вычисления для критерия: в ячейке F2 вычисляется значение критериальной статистики по формуле

$$=\text{КОРЕНЬ}(\text{C3}) * (\text{C4} - \text{C1}) / \text{C5},$$

в ячейках E5:G5 — критические значения по формулам соответственно

$$=\text{СТЮДРАСПОБР}(\text{\$C\$2}/2; \text{\$C\$3} - 1)$$

для случая равенства и

$$=\text{СТЮДРАСПОБР}(\text{\$C\$2}; \text{\$C\$3} - 1)$$

для случая неравенств (в ячейках F5 и G5 одинаковые формулы). В ячейках E7:G7 проверяются условия выполнения критерия (формулы показаны на рис. 11.1).

F2		=КОРЕНЬ(C3)*(C4-C1)/C5					
	A	B	C	D	E	F	G
1	Выборка m0=		-0,5		Критериальная статистика		
2	-1,60046	Уровень значимости	0,05			-2,169829854	
3	-3,55537	Объем выборки	50		Равенство	Неравенство б)	Неравенство в)
4	-0,51149	Среднее	-1,214		Квантиль tв	Квантиль tв	Квантиль tн
5	1,552947	Ст. отклонение	2,3268		2,312372089	2,009574018	2,009574018
6	1,3967				Гипотеза		
7	2,466266				принимается	принимается	отклоняется
8	-5,36718	=ЕСЛИ(ABS(\$F\$2)<E5,"принимается","отклоняется")					
9	-1,46836	=ЕСЛИ(\$F\$2<F5,"принимается","отклоняется")					
10	1,190045	=ЕСЛИ(\$F\$2>G5,"принимается","отклоняется")					
11	-3,1734						
12	-2,38041						

Рис. 11.1. Критерий проверки значения математического ожидания

Как видно на рис. 11.1, для $m_0 = -0,5$ нулевые гипотезы в случае равенства и неравенства б) принимаются (напомним, что истинное значение математического ожидания равно -1), а нулевая гипотеза для неравенства в) отклоняется. Посмотрим, как среагирует критерий, если положить $m_0 = 0$. Результат применения критерия для этого случая показан на рис. 11.2. Здесь гипотеза о равенстве отвергается.

	A	B	C	D	E	F	G
1	Выборка m0=		0		Критериальная статистика		
2	-1,60046	Уровень значимости	0,05			-3,689304685	
3	-3,55537	Объем выборки	50		Равенство	Неравенство б)	Неравенство в)
4	-0,51149	Среднее	-1,214		Квантиль tв	Квантиль tв	Квантиль tн
5	1,552947	Ст. отклонение	2,3268		2,312372089	2,009574018	2,009574018
6	1,3967				Гипотеза		
7	2,466266				отклоняется	принимается	отклоняется
8	-5,36718						
9	-1,46836						
10	1,190045						
11	-3,1734						
12	-2,38041						

Рис. 11.2. Сравнение математического ожидания с нулевым значением

11.1.2. Критерий проверки значения дисперсии нормальной совокупности

Статистическая модель. Выборка x_1, x_2, \dots, x_n получена из генеральной совокупности с нормальным законом распределения с неизвестными математическим ожиданием μ и дисперсией σ^2 .

Гипотезы

а) Равенство

$$H_0: \sigma^2 = \sigma_0^2$$

$$H_1: \sigma^2 \neq \sigma_0^2$$

б) Неравенство

$$H_0: \sigma^2 \leq \sigma_0^2$$

$$H_1: \sigma^2 > \sigma_0^2$$

в) Неравенство

$$H_0: \sigma^2 \geq \sigma_0^2$$

$$H_1: \sigma^2 < \sigma_0^2$$

Здесь σ_0^2 — заданное число. Задан уровень значимости α .

Вычисления. Вычисляется критериальная статистика $T = \frac{(n-1)S_n^2}{\sigma_0^2}$, где

$$S_n^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2.$$

Построение критической области. При условии истинности гипотезы H_0 статистика T имеет распределение χ^2 с $(n-1)$ степенью свободы.

Случай а). Вычисляются критические значения как квантили t_α порядка $\alpha/2$ и $t_{1-\alpha/2}$ порядка $1 - \alpha/2$ распределения χ^2 с $(n-1)$ степенью свободы. Гипотеза H_0 принимается, если выполняется неравенство $t_\alpha \leq T \leq t_{1-\alpha/2}$, иначе гипотеза H_0 отклоняется.

Случай б). Вычисляется критическое значение как квантиль t_α порядка $1 - \alpha$ распределения χ^2 с $(n-1)$ степенью свободы. Гипотеза H_0 принимается, если $T \leq t_\alpha$.

Случай в). Вычисляется критическое значение как квантиль t_α порядка α распределения χ^2 с $(n-1)$ степенью свободы. Гипотеза H_0 принимается, если $t_\alpha \leq T$.

Комментарий. Критерий не устойчив, если не выполняется условие нормальности распределения генеральной совокупности.

Практическая реализация

На рис. 11.3 показан рабочий лист Excel со всеми формулами, необходимыми для реализации критерия. Выборка имеет нормальное распределение с математическим ожиданием 1 и дисперсией 4. Значение σ_0^2 задано в ячейке C1, а уровень значимости — в ячейке C2. Отметим, что выборочная дисперсия в ячейке C5 вычисляется с помощью функции ДИСПР (см. раздел 4.5.2).

C4		A = СРЗНАЧ(Выборка)				
A		B		Критериальная статистика		
1	Выборка	Сигма0=	4			
2	0,399536	Уровень значимости	0,01	115,5134165 =(C3-1)*C5/C1		
3	-1,55537	Объем выборки	100	Равенство	Неравенство б)	Неравенство в)
4	1,488515	Среднее	0,919	Квантиль t_α	Квантиль t_α	Квантиль t_α
5	3,552947	Выборочная дисперсия	4,6672	138,986918	134,6414896	69,22985585
6	3,3967	=ДИСПР(Выборка)		Квантиль t_α	=ХИ2ОБР(\$C\$2,\$C\$3-1)	
7	4,466266	=ХИ2ОБР(\$C\$2/2,\$C\$3-1)		66,50990655	=ХИ2ОБР(1-\$C\$2,\$C\$3-1)	
8	-3,36718	=ХИ2ОБР(1-\$C\$2/2,\$C\$3-1)		Гипотеза		
9	0,531638			принимается	принимается	принимается
10	3,190045	=ЕСЛИ(И(\$F\$2<E5,\$F\$2>E7),				
11	-1,1734	"принимается","отклоняется")				
12	-0,38041	=ЕСЛИ(\$F\$2<F5,"принимается","отклоняется")				
13	-2,38086	=ЕСЛИ(\$F\$2>G5,"принимается","отклоняется")				
14	-2,69382					
15	-0,95526					
16	-0,54701					

Рис. 11.3. Критерий проверки значения дисперсии

Как видно на рис. 11.3, для случая $\sigma_0^2 = 4$ принимаются все три нулевые гипотезы. На рис. 11.4 показаны результаты вычисления критерия для $\sigma_0^2 = 2$. Здесь нулевые гипотезы равенства и неравенства б) отклоняются, а гипотеза $H_0: \sigma^2 \geq \sigma_0^2$ принимается. Выполнение критерия для случая $\sigma_0^2 = 7$ показано на рис. 11.5.

	A	B	C	D	E	F	G
1	Выборка	Сигма0=	2	Критериальная статистика			
2	0,399536	Уровень значимости	0,01		231,026833		
3	-1,55537	Объем выборки	100	Равенство	Неравенство б)	Неравенство в)	
4	1,488515	Среднее	0,919	Квантиль tв	Квантиль tв	Квантиль tн	
5	3,552947	Выборочная	4,6672	138,986918	134,6414896	69,22985585	
6	3,3967	дисперсия		Квантиль tн			
7	4,466266			66,50990655			
8	-3,36718			Гипотеза			
9	0,531638			отклоняется	отклоняется	принимается	
10	3,190045						
11	-1,1734						
12	-0,38041						

Рис. 11.4. Проверка значения дисперсии для $\sigma_0^2 = 2$

	A	B	C	D	E	F	G
1	Выборка	Сигма0=	7	Критериальная статистика			
2	0,399536	Уровень значимости	0,01		66,00766656		
3	-1,55537	Объем выборки	100	Равенство	Неравенство б)	Неравенство в)	
4	1,488515	Среднее	0,919	Квантиль tв	Квантиль tв	Квантиль tн	
5	3,552947	Выборочная	4,6672	138,986918	134,6414896	69,22985585	
6	3,3967	дисперсия		Квантиль tн			
7	4,466266			66,50990655			
8	-3,36718			Гипотеза			
9	0,531638			отклоняется	принимается	отклоняется	
10	3,190045						
11	-1,1734						
12	-0,38041						

Рис. 11.5. Проверка значения дисперсии для $\sigma_0^2 = 7$

11.2. Проверка гипотезы о значении параметра показательного распределения

Показательное (экспоненциальное) распределение определяется параметром λ (см. раздел 1.5.3), при этом для случайной величины X , подчиняющейся этому распределению, $MX = 1/\lambda$, $DX = 1/\lambda^2$. Для этого распределения обычно строятся критерии оценки не параметра λ , а обратная к нему величина $\theta = 1/\lambda$, поскольку $MX = \theta$. Построение критериев для параметра θ основано на том, что случай-

ная величина $2 \sum_{i=1}^n x_i / \theta$, где x_i — выборочные значения, имеющие показательное распределение с параметром θ , не зависит от параметра θ и имеет распределение χ^2 с $2n$ степенями свободы.

Статистическая модель. Генеральная совокупность имеет показательное распределение с параметром θ .

Гипотезы

а) Равенство

$$H_0: \theta = \theta_0$$

$$H_1: \theta \neq \theta_0$$

б) Неравенство

$$H_0: \theta \leq \theta_0$$

$$H_1: \theta > \theta_0$$

в) Неравенство

$$H_0: \theta \geq \theta_0$$

$$H_1: \theta < \theta_0$$

Здесь θ_0 — заданное число. Задан уровень значимости α .

Вычисления. Вычисляется критерияльная статистика $T = \frac{2n\bar{x}}{\theta_0}$, где $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$.

Построение критической области. При условии истинности гипотезы H_0 статистика T имеет распределение χ^2 с $2n$ степенью свободы.

Случай а). Вычисляются критические значения t_n как квантиль порядка $\alpha/2$ распределения χ^2 с $2n$ степенью свободы и t_n как квантиль порядка $1 - \alpha/2$ того же распределения. Гипотеза H_0 принимается, если выполняется неравенство $t_n \leq T \leq t_n$, иначе гипотеза H_0 отклоняется.

Случай б). Вычисляется квантиль t_n порядка $1 - \alpha$ распределения χ^2 с $2n$ степенью свободы. Гипотеза H_0 принимается, если $T \leq t_n$.

Случай в). Вычисляется квантиль t_n порядка α распределения χ^2 с $2n$ степенью свободы. Гипотеза H_0 принимается, если $t_n \leq T$.

Комментарий. Критерий практически тождествен методу построения доверительного интервала для параметра θ (см. раздел 10.5).

Практическая реализация

На рис. 11.6 показана выборка (столбец А), имеющая показательное распределение с параметром $\lambda = 0,5$ (или $\theta = 2$). Выборка построена с помощью формулы массива $\{=\text{ГАММАОБР}(\text{СЛЧИС}();1;2)\}$. В Excel нет функции, обратной к функции распределения показательного закона, но поскольку это распределение является частным случаем гамма-распределения при $\alpha = 1$, можно воспользоваться функцией ГАММАОБР (см. раздел 4.7.3), если положить в ней второй аргумент равным 1. Отметим, что третий аргумент в этой функции задает параметр θ , а не λ . Все формулы, необходимые для построения критерия, показаны на рис. 11.6.

	C6		=CPЗНАЧ(Выборка)				
	A	B	C	D	E	F	G
1	Выборка $\theta_0=$	1,7	Статистика	125,0804376	=2*C5*C6/C1		
2	3,716132	Уровень значимости	Критические значения для гипотез				
3	0,986445	0,1	Равенство	Неравенство б)	Неравенство в)		
4	6,578994	Объем выборки	124,3421013	118,4980018	82,3581269		
5	0,339693	50	77,92944231	=ХИ2ОБР(1-С\$3;2*С\$5)			
6	0,728646	Среднее 2,126367	=ХИ2ОБР(С\$3;2*С\$5)				
7	4,41567	Гипотеза					
8	1,42214	=ХИ2ОБР(1-С3/2;2*С5)	отклоняется	отклоняется	принимается		
9	5,520742						
10	3,860405				=ЕСЛИ(F1>G4,"принимается","отклоняется")		
11	11,77425				=ЕСЛИ(F1<F4,"принимается","отклоняется")		
12	6,446062	=ЕСЛИ(И(F1<E4,E5<F1),"принимается","отклоняется")					
13	13,29812						
14	1,353292						

Рис. 11.6. Критерии проверки значения параметра показательного распределения

11.3. Проверка гипотезы о значении параметра биномиального распределения

Построение критерия проверки гипотезы о значении параметра биномиального распределения, как и при построении доверительного интервала для биномиальной вероятности (см. раздел 10.8), можно рассмотреть отдельно для случаев, когда выборка содержит наблюдения за одним экспериментом и когда выборка содержит результаты нескольких независимых экспериментов. Однако, поскольку критерии в обоих случаях с вычислительной точки зрения практически не отличаются, рассмотрим только случай, когда выборка содержит наблюдения за одним экспериментом.

Статистическая модель. Выборка x_1, x_2, \dots, x_n является результатом наблюдения за одним экспериментом, состоящим из n одинаковых испытаний, в каждом из которых с вероятностью p может произойти исход "1" и с вероятностью $(1 - p)$ — исход "0". Здесь x_i равно 1, если в i -м испытании произошел исход "1", и 0 в противном случае.

Гипотезы

а) Равенство

$$H_0: p = p_0$$

$$H_1: p \neq p_0$$

б) Неравенство

$$H_0: p \leq p_0$$

$$H_1: p > p_0$$

в) Неравенство

$$H_0: p \geq p_0$$

$$H_1: p < p_0$$

Здесь p_0 — заданное число. Задан уровень значимости α .

Как указывалось в разделе 10.8, несмещенной и эффективной оценкой для вероятности p будет статистика $\hat{p} = r/n$, где r — количество исходов "1". Случайная величина r имеет биномиальное распределение с параметрами n и p . Распределение статистики \hat{p} асимптотически нормально с параметрами $m = p$ и $\sigma^2 = p(1 - p)/n$.

Так же, как и доверительные интервалы для вероятности p , критерии проверки значения этой вероятности строятся или на основе биномиального распределения, которое имеет случайная величина r , или на основе асимптотической нормальности распределения статистики \hat{p} .

11.3.1. Использование биномиального распределения

Сразу отметим, что здесь биномиальное распределение заменяется F -распределением по известному соотношению $P(X \leq k) = F_{n-k, k+1}(1 - p)$, где X — случайная величина, имеющая биномиальное распределение с параметрами n и p , а $F_{n-k, k+1}$ — функция F -распределения с соответствующими параметрами. Это же соотношение использовалось в разделе 10.8.1 при построении доверительного интервала для вероятности p .

Вычисления. Вычисляются критерийные статистики $Y = \frac{r}{n+1-r} \frac{1-p_0}{p_0}$,

$Z = \frac{n-r}{r+1} \frac{p_0}{1-p_0}$. Обе статистики применяются для критерия равенства; для критерия неравенства б) используется статистика Y , для критерия неравенства в) — статистика Z .

Построение критической области. При условии истинности нулевых гипотез величина r имеет биномиальное распределение с параметрами n и p . Для вычисления квантилей этого распределения, которые необходимы для построения критической области, используется F -распределение с соответствующими значениями степеней свободы.

Случай а). Вычисляются критические значения t_1 как квантиль порядка $1 - \alpha/2$ F -распределения с $2(n + 1 - r)$ и $2r$ степенями свободы и t_2 как квантиль порядка $1 - \alpha/2$ F -распределения с $2(r + 1)$ и $2(n - r)$ степенями свободы. Гипотеза H_0 принимается, если выполняются неравенства $Y \leq t_1$ и $Z \leq t_2$. Если хотя бы одно из этих неравенств не выполняется, гипотеза H_0 отклоняется.

Случай б). Вычисляется критическое значение t_α как квантиль порядка $1 - \alpha$ F -распределения с $2(n + 1 - r)$ и $2r$ степенями свободы. Гипотеза H_0 принимается, если $Y \leq t_\alpha$.

Случай в). Вычисляется критическое значение t_α как квантиль порядка α F -распределения с $2(r + 1)$ и $2(n - r)$ степенями свободы. Гипотеза H_0 принимается, если $t_\alpha \leq Z$.

Комментарий. Данный критерий, в основном, применяется к выборкам малого объема. Для выборок большого объема чаще применяется асимптотический критерий, основанный на аппроксимации биномиального распределения нормальным (см. следующий раздел).

Практическая реализация

На рис. 11.7 в столбце А показана выборка, содержащая 30 наблюдений за экспериментом, где с вероятностью 0,4 происходит событие "1". Эта выборка получена с помощью средства пакета анализа Генерация случайных чисел. Все формулы Excel, необходимые для построения критерия, также показаны на рис. 11.7. В ячейках С4:С7 для упрощения формул, вычисляющих квантили, подсчитаны соответствующие значения степеней свободы: $2(n + 1 - r)$, $2r$, $2(r + 1)$ и $2(n - r)$.

C2		=СУММ(Выборка)							
A	B	C	D	E	F	G	H	I	
1	Выборка n=	30	p=	0,5					
2	0 r=	14	Уровень значимости		0,05				
3	1 p=	0,47	Y=	0,823529412	= (C2/(C1+1-C2))^(1-E1)/E1				
4	0 k1=	34	Z=	1,066666667	= ((C1-C2)/(C2+1))^E1/(1-E1)				
5	1 k2=	28	Критические значения для гипотез						
6	0 k3=	30	Равенство	Неравенство б)	Неравенство в)				
7	0 k4=	32	2,082174433	1,846245823	0,546787504	=FРАСПОБР(1-\$F\$2,C6,C7)			
8	0		2,040835057	=FРАСПОБР(\$F\$2,\$C\$4,\$C\$5)					
9	1		Гипотеза						
10	1		принимается		отклоняется				
11	1		=ЕСЛИ(E4<F7,"принимается","отклоняется")						
12	0		=FРАСПОБР(\$F\$2/2,C6,C7)						
13	1		=FРАСПОБР(\$F\$2/2,\$C\$4,\$C\$5)						
14	0		=ЕСЛИ(E3<E7,"принимается","отклоняется")						
15	1		=ЕСЛИ(И(E3<D7,E4<D8),"принимается","отклоняется")						
16	1								

Рис. 11.7. Критерий проверки значения вероятности p

ки значения медианы, но и для других целей, например как оценка центра местоположения распределения, если по каким-либо причинам нельзя для этих целей использовать математическое ожидание. (Для некоторых распределений математическое ожидание может просто не существовать, как, например, у распределении Коши.) Кроме того, описанные ниже критерии являются свободными от распределения, т.е. непараметрическими. Поэтому их можно использовать "без оглядки" на исходное распределение выборки, тип которого часто трудно определить. Для симметричных распределений медиана и математическое ожидание совпадают, поэтому данные критерии также можно использовать для проверки значений математических ожиданий таких распределений. Но если все-таки известен тип распределения, то в последнем случае надежнее применять критерии, использующие информацию о типе распределения.

11.4.1. Критерий знаков

Статистическая модель. Выборочные значения x_1, x_2, \dots, x_n независимы и взяты из одной генеральной совокупности. Значение медианы m неизвестно.

Гипотезы

а) Равенство

б) Неравенство

в) Неравенство

$H_0: m = m_0$

$H_0: m \leq m_0$

$H_0: m \geq m_0$

$H_1: m \neq m_0$

$H_1: m > m_0$

$H_1: m < m_0$

Здесь m_0 — заданное число. Задан уровень значимости α .

В этом критерии в качестве критериальной статистики используется подсчитанное число R выборочных значений, которые больше m_0 . Если справедливы нулевые гипотезы, то случайная величина R имеет биномиальное распределение с параметрами n и $p = 0,5$. Как и в критериях проверки биномиальных вероятностей, здесь можно построить или точный критерий, основанный на биномиальном распределении величины R , либо асимптотический, использующий аппроксимацию биномиального распределения нормальным. Рассмотрим сначала точный критерий.

Точный критерий знаков

Вычисления. Вычисляется критериальная статистика R , равная количеству выборочных значений, которые по величине больше m_0 . Дополнительно вычисляются статистики $Y = \frac{R}{n+1-R}$, $Z = \frac{n-R}{R+1}$. Обе статистики применяются для

критерия равенства; для критерия неравенства б) используется статистика Y , для критерия неравенства в) — статистика Z .

Построение критической области. При условии истинности нулевых гипотез статистика R имеет биномиальное распределение с параметрами n и $0,5$. Однако для получения критических значений, которые основаны на квантилях биномиального распределения, как и в критерии о значении биномиальной вероятности (раздел 11.3.1), используется F -распределение.

Случай а). Вычисляются критические значения t_1 как квантиль порядка $1 - \alpha/2$ F -распределения с $2(n+1-R)$ и $2R$ степенями свободы и t_2 как квантиль порядка $1 - \alpha/2$ F -распределения с $2(R+1)$ и $2(n-R)$ степенями свободы. Гипотеза H_0 принимается, если выполняются неравенства $Y \leq t_1$ и $Z \leq t_2$. Если хотя бы одно из этих неравенств не выполняется, гипотеза H_0 отклоняется.

Случай б). Вычисляется критическое значение t_n как квантиль порядка $1 - \alpha$ F-распределения с $2(n + 1 - R)$ и $2R$ степенями свободы. Гипотеза H_0 принимается, если $Y \leq t_n$.

Случай в). Вычисляется критическое значение t_n как квантиль порядка α F-распределения с $2(R + 1)$ и $2(n - R)$ степенями свободы. Гипотеза H_0 принимается, если $t_n \leq Z$.

Комментарии

1. Если какое-либо выборочное значение равно m_0 , то оно не учитывается, а значение n уменьшается на 1.
2. Данный критерий, в основном, применяется к выборкам малого объема. Для выборок большого объема чаще применяется асимптотический критерий, основанный на аппроксимации биномиального распределения нормальным (см. следующий раздел).

Практическая реализация

На рис. 11.9 в столбце А показана выборка, содержащая 30 значений, равномерно распределенных на интервале $[0, 10]$ (таким образом, истинное значение медианы равно 5). Все формулы Excel, необходимые для построения критерия, также показаны на рис. 11.9. Значение статистики R вычисляется с помощью формулы массива

{=СЧЕТ(ЕСЛИ(Выборка>Е1;Выборка;""))}.

Работа подобных формул описана в разделе 6.1.4. Поскольку для непрерывных распределений вероятность того, что случайная величина примет какое-либо конкретное значение, равна нулю, то здесь проверка на совпадение выборочного значения и m_0 не выполняется. Добавить подобную проверку в расчетные формулы несложно, но возникает необходимость в некоторых промежуточных вычислениях. В ячейках C4:C7 для упрощения формул, вычисляющих квантили, отдельно подсчитаны соответствующие значения степеней свободы: $2(n + 1 - r)$, $2r$, $2(r + 1)$ и $2(n - r)$.

C2		{=СЧЕТ(ЕСЛИ(Выборка>Е1;Выборка;""))}							
1	Выборка n =	30	m ₀ =	6					
2	7,86443	R =	14	Уровень значимости	0,05				
3	8,69812	Y =		0,823529412	=C2/(C1+1-C2)				
4	6,33307	k1 =	34	Z =	1,066666667	= (C1-C2)/(C2+1)			
5	9,2283	k2 =	28	Критические значения для гипотез					
6	3,23518	k3 =	30	Равенство Неравенство б) Неравенство в)					
7	4,77486	k4 =	32	2,082174433	1,846245823	0,546787504	=FРАСПОБР(1-\$F\$2,C6,C7)		
8	5,98045			2,040835057	=FРАСПОБР(\$F\$2,\$C\$4,\$C\$5)				
9	6,14012			Гипотеза					
10	4,26364			принимается, принимается, отклоняется					
11	5,29475			=ЕСЛИ(E4<F7,"принимается","отклоняется")					
12	1,87057	=FРАСПОБР(\$F\$2/2,C6,C7)		=ЕСЛИ(E3<E7,"принимается","отклоняется")					
13	2,95179	=FРАСПОБР(\$F\$2/2,\$C\$4,\$C\$5)		=ЕСЛИ(И(E3<D7,E4<D8),					
14	1,62359			"принимается","отклоняется")					
15	7,37216								
16	4,11789								

Рис. 11.9. Точный критерий знаков

Асимптотический критерий знаков

Вычисления. Вычисляется статистика R , равная количеству выборочных значений, которые по величине больше m_0 . Далее вычисляется критериальная статистика $T = \frac{2R - n}{\sqrt{n}}$.

Построение критической области. При условии истинности нулевых гипотез статистика T имеет асимптотически стандартное нормальное распределение.

Случай а). Вычисляются критические значения t как квантиль порядка $1 - \alpha/2$ стандартного нормального распределения. Гипотеза H_0 принимается, если $|T| \leq t$. В противном случае гипотеза H_0 отклоняется.

Случай б). Вычисляется критическое значение t_1 как квантиль порядка $1 - \alpha$ стандартного нормального распределения. Гипотеза H_0 принимается, если $T \leq t_1$.

Случай в). Вычисляется критическое значение t_2 как квантиль порядка α стандартного нормального распределения. Гипотеза H_0 принимается, если $t_2 \leq T$.

Комментарии

1. Если какое-либо выборочное значение равно m_0 , то оно не учитывается, а значение n уменьшается на 1.

2. Данный критерий является приближенным и применяется, в основном, к выборкам большого объема ($n > 50$). Если $n \leq 50$, рекомендуется использо-

вать "исправленную" статистику T вида $T = \frac{2R - n + 1}{\sqrt{n}}$, которая также име-

ет асимптотически стандартное нормальное распределение. Для малых выборок ($n \leq 20$) применяется критерий, описанный в предыдущем разделе.

Практическая реализация данного критерия, если подсчитано значение R , не вызывает трудностей. Формула для вычисления значения R показана в предыдущем разделе.

11.4.2. Критерий знаковых рангов Уилкоксона

Считается, что критерий знаков, описанный в предыдущем разделе, не учитывает значительную часть информации, содержащейся в выборке. Критерий знаковых рангов Уилкоксона не только считает количество отрицательных или положительных разностей $x_i - m_0$, но и учитывает через значения рангов относительные размеры этих разностей.

Статистическая модель. Выборочные значения x_1, x_2, \dots, x_n независимы и взяты из одной генеральной совокупности. Значение медианы m неизвестно.

Гипотезы

а) Равенство

$$H_0: m = m_0$$

$$H_1: m \neq m_0$$

б) Неравенство

$$H_0: m \leq m_0$$

$$H_1: m > m_0$$

в) Неравенство

$$H_0: m \geq m_0$$

$$H_1: m < m_0$$

Здесь m_0 — заданное число. Задан уровень значимости α .

1. Вычисляются ранги r_i величин $|x_i - m_0|$. Значения, для которых $x_i - m_0 = 0$, игнорируются.
2. Вычисляются критериальные статистики V_+ и V_- , равные сумме рангов r_i положительных разностей $x_i - m_0$ и сумме рангов отрицательных разностей $x_i - m_0$ соответственно.

Построение критической области. При условии истинности нулевых гипотез статистики V_+ и V_- распределены одинаково. Для вычисления критических значений используются квантили специального распределения, которое имеют величины V_+ и V_- [9]. Эти величины также имеют асимптотически нормальное распределение с математическим ожиданием $n(n+1)/2$ и дисперсией $n(n+1)(2n+1)/24$. Поскольку статистики V_+ и V_- имеют одинаковые распределения, то для проверки гипотез используется одна из этих статистик. Возьмем в качестве критериальной статистики величину V_+ и используем ее асимптотическую нормальность. Для удобства вычислим величину $T = \frac{V_+ - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$,

которая имеет асимптотически стандартное нормальное распределение.

Случай а). Вычисляется критическое значение t как квантиль порядка $1 - \alpha/2$ стандартного нормального распределения. Гипотеза H_0 принимается, если $|T| \leq t$. В противном случае гипотеза H_0 отклоняется.

Случай б). Вычисляется критическое значение t_1 как квантиль порядка $1 - \alpha$ стандартного нормального распределения. Гипотеза H_0 принимается, если $T \leq t_1$.

Случай в). Вычисляется критическое значение t_2 как квантиль порядка α стандартного нормального распределения. Гипотеза H_0 принимается, если $t_2 \leq T$.

Комментарии

1. Описанный критерий с использованием нормального распределения применяется для выборок, объем которых больше 20. Для малых выборок необходимо использовать точное распределение статистик V_+ и V_- .
2. В [18, с. 124] приводится другая аппроксимация распределения величин V_+ и V_- , построенная на основе распределения Стьюдента.

Практическая реализация

На рис. 11.10 в столбце А показана выборка из 30 значений, имеющих равномерное распределение на интервале $[0, 10]$, и основные формулы, необходимые для реализации критерия. К сожалению, в Excel не удастся подсчитать значение V_+ с помощью одной формулы только по выборочным значениям. Для вычисления этого значения пришлось отдельно подсчитать абсолютные величины разностей $x_i - m_0$ (столбец В, формула массива $\{=ABS(Выборка-G1)\}$, диапазон ячеек с выборочными значениями назван Выборка). В столбце С подсчитаны ранги абсолютных величин разностей $x_i - m_0$ по формуле массива $\{=РАНГ(Разности;Разности;1)\}$ (здесь диапазон ячеек со значениями в столбце В назван Разности, функция РАНГ описана в разделе 4.2.5). После этих предварительных вычислений значение V_+ вычисляется в ячейке E2 с помощью формулы массива $\{=СУММ(ЕСЛИ(Выборка>G1;Ранги;""))\}$ (в ячейке G1 содержится значение m_0 , диапазон ячеек, содержащий ранги, назван Ранги). Остальные формулы показаны на рис. 11.10.

E2 {=СУММ(ЕСЛИ(Выборка>G1;Ранги;""))}									
A	B	C	D	E	F	G	H	I	
1	Выборка	Разности	Ранги	n = 30	m = 6				
2	3,85947	2,14053	17	V+	195	Уровень значимости	0,05		
3	5,96871	0,031287	1			Критериальная статистика T			
4	0,8469	5,153103	28			-0,771312726	=(E2-E1*(E1+1)/4)/КОРЕНЬ(E1*(E1+1))		
5	9,48607	3,486074	24			Критические значения для гипотез $\sqrt{(2*E1+1)/24}$			
6	4,84207	1,157925	12			Равенство	Неравенство б)	Неравенство в)	
7	6,12907	0,129071	2			1,959962787	1,644853476	-1,644853476	
8	5,7161	0,283898	5			=НОРМСТОБР(1-\$H\$2)			
9	9,95087	3,950869	26			=НОРМСТОБР(1-\$H\$2/2)			=НОРМСТОБР(\$H\$2)
10	9,59726	3,597258	25			Гипотеза			
11	6,78628	0,786279	9			принимается	принимается	отклоняется	
12	0,15915	5,840849	30						
13	5,13851	0,861495	10			=ЕСЛИ(G3<G7,"принимается","отклоняется")			
14	7,37441	1,374414	13			=ЕСЛИ(G4<H7,"принимается","отклоняется")			
15	6,22989	0,229888	4			=ЕСЛИ(ABS(G4)<F7,"принимается","отклоняется")			
16	5,84407	0,155929	3						
17	6,7388	0,738795	7						

Рис. 11.10. Реализация критерия знаковых рангов Уилкоксона

Сравнение одномерных выборок

Если имеется несколько одномерных выборок, то, прежде чем приступить к их статистическому анализу, обычно ставят два следующих общих вопроса.

- Имеют ли эти выборки одинаковые распределения, или, другими словами, получены ли они из одной генеральной совокупности?
- Имеют ли значения выборочных параметров значимые различия или их можно считать равными?

На первый вопрос помогают ответить методы сравнения выборочных распределений, описанные в разделе 12.1. На второй вопрос можно ответить двумя способами: путем построения доверительных интервалов для разностей или отношений сравниваемых параметров либо с помощью критериев проверки гипотез о значимых разностях или отношений этих параметров. Построение доверительных интервалов показано в разделе 12.2, а критерии проверки гипотез — в разделе 12.3.

12.1. Сравнение выборочных распределений

Для сравнения выборочных распределений разработано много критериев (их часто называют *критериями однородности*), имеющих различные теоретические основы. Эти критерии, как правило, непараметрические, поскольку, если известен класс распределений, которому подчиняются выборочные значения, в этом случае ставится задача сравнения не самих распределений, а их параметров, и эта задача решается иными методами.

В принципе, описанные ниже критерии можно применять и для сравнения параметров распределений, если априори предположить, что выборочные распределения принадлежат одному классу распределений и необходимо сравнить значения одного параметра распределения. В этом случае отклонение нулевой гипотезы, состоящей в том, что выборочные распределения совпадают, говорит о том, что значения данного параметра различны. Однако с помощью этих критериев невозможно оценить степень различия значений параметров.

Рассмотрим несколько критериев, начиная с наиболее простых (и менее точных). Обращаем внимание, что большинство описанных критериев рассчитано на непрерывные распределения либо требуют некоторых модификаций для работы с дискретными распределениями. Для сравнения дискретных распределений рекомендуем сразу обратиться к критерию хи-квадрат (раздел 12.5), если нет каких-либо "противопоказаний" или если с помощью этих критериев вы не проверите различие в значениях параметров распределений, как сказано выше.

12.1.1. Непараметрический критерий медианы

Этот критерий является модификацией критерия знаков для проверки гипотез о значении медианы (см. раздел 11.4.1), обобщенный для случая нескольких выборок.

Статистическая модель. Имеется k одномерных независимых выборок объемом соответственно n_1, n_2, \dots, n_k .

Гипотезы

H_0 : все k выборок имеют одинаковые распределения;

H_1 : нулевая гипотеза неверна.

Задан уровень значимости α .

Вычисления

1. Все k выборок объединяются в единую выборку, и по объединенной выборке

вычисляется выборочная медиана m следующим образом (далее $n = \sum_{i=1}^k n_i$).

а) Для непрерывных распределений по объединенной выборке строится вариационный ряд $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$. Выборочная медиана $m = x_{(k+1)}$, если $n = 2k + 1$, и $m = (x_{(k)} + x_{(k+1)})/2$, если $n = 2k$.

б) Для дискретных распределений по объединенной выборке рассчитывается частотная таблица, которая сортируется по возрастанию значений. Вычисляется значение x_m , которому соответствует накопленная частота F_m , меньшая $n/2$, и следующее по величине значение x_{m+1} , которому соответствует накопленная частота F_{m+1} , большая или равная $n/2$. Тогда медиана m вычисляется по формуле $M = x_m + (x_{m+1} - x_m) \frac{n/2 - F_m}{F_{m+1} - F_m}$.

2. Для каждой i -й выборки подсчитывается число значений R_i , превосходящих m . Если есть одиночные выборочные значения, совпадающие со значением m , то эти значения исключаются из подсчетов, а значение объема соответствующей выборки уменьшается на единицу. Для дискретных распределений вероятно совпадение с выборочной медианой m сразу нескольких значений в одной выборке. В этом случае, если таких значений четное число, половина из них считается меньшими m , а половина большими m . Если же таких значений нечетное число, то отбрасывается одно значение (объем выборки также уменьшается на единицу), а остальные делятся пополам и считается, что одна половина больше m , а другая — меньше m . Обычно для удобства вычислений составляется таблица следующего вида.

	Число значений, больших m	Число значений, меньших m	Всего
Выборка 1	R_1	$n_1 - R_1$	n_1
Выборка 2	R_2	$n_2 - R_2$	n_2
...
Выборка k	R_k	$n_k - R_k$	n_k
Всего	$\sum_{i=1}^k R_i$	$n - \sum_{i=1}^k R_i$	$n = \sum_{i=1}^k n_i$

3. Вычисляется критериальная статистика $G = \frac{\sum (Y_i - D_i) R_i}{n/2} - p$.

Построение критической области. При условии истинности нулевой гипотезы статистика T имеет распределение χ^2 с $(k - 1)$ степенью свободы.

Вычисляются критические значения t_α как квантиль порядка $\alpha/2$ распределения χ^2 с $(k - 1)$ степенью свободы и $t_{1-\alpha/2}$ как квантиль порядка $1 - \alpha/2$ того же распределения. Гипотеза H_0 принимается, если выполняется неравенство $t_\alpha \leq T \leq t_{1-\alpha/2}$, иначе гипотеза H_0 отклоняется.

Комментарий. Как и критерий Уилкоксона-Манна-Уитни (см. следующий раздел), этот критерий скорее “улавливает” различия в положении медиан выборок, а не различия в форме распределений. Поэтому, с одной стороны, ему следует, по возможности, предпочесть более надежные критерии, а с другой стороны, его можно использовать как критерий совпадения “средних значений” выборок.

Практическую реализацию покажем отдельно для непрерывных и дискретных распределений.

Реализация критерия для непрерывных распределений

Сначала покажем, как подсчитать значение выборочной медианы. На рис. 12.1 показаны три выборки объемом соответственно 20, 30 и 40 значений. Все выборки имеют нормальные распределения, причем первые две выборки — стандартное, а третья — нормальное распределение с единичной дисперсией и математическим ожиданием, равным 1. Выборки получены с помощью формул массивов

{=НОРМСТОБР(СЛЧИС())} и {=НОРМОБР(СЛЧИС();1;1)}.

D2 (=РАНГ(A2:C41;A2:C41*1))									
	A	B	C	D	E	F	G	H	I
	Выборка1	Выборка2	Выборка3	Ранг1	Ранг2	Ранг3	Объемы		
2	2,733263	0,13507	0,4417877	89	33	44	Выборка1	20	=СЧЕТ(Выборка1)
3	-0,209671	-0,487504	-0,320153	22	16	18	Выборка2	30	=СЧЕТ(Выборка2)
4	-0,849079	0,517294	0,7653465	9	25	52	Выборка3	40	=СЧЕТ(Выборка3)
5	1,874827	0,255092	-0,183674	80	40	23	Всего	90	=СУММ(H2:H4)
6	0,182413	-0,179108	1,6814509	37	24	74	Ранг1	45	=ЦЕЛОЕ((H5+1)/2)
7	-0,067692	0,518611	1,3474805	27	24	68	Ранг2	46	=ЕСЛИ(ЕЧЕТН(H5),H6+1,H6)
8	0,17736	-1,207186	0,2793753	36	6	41	Значение1	0,51729	
9	2,145176	0,426176	1,6613293	87	43	73	Значение2	0,51861	
10	-0,314232	-0,590907	1,5477334	20	14	71	Медиана	0,51795	=(H8+H9)/2
11	0,026103	-0,024356	-0,318949	30	28	19			
12	1,884378	-0,666326	0,7199918	81	13	50			=БИЗВЛЕЧЬ(A1:F41;ЕСЛИ(ЕНД(ПОИСКПОЗ(H6;Ранг1;0));ЕСЛИ(ЕНД(ПОИСКПОЗ(H6;Ранг2;0));3;2);1);J1:L4)
13	1,868212	-2,325264	1,8604737	79	1	78			
14	-1,44609	0,538928	0,835437	4	47	54			
15	0,558783	-0,447681	2,0357333	48	17	86			=БИЗВЛЕЧЬ(A1:F41;ЕСЛИ(ЕНД(ПОИСКПОЗ(H7;Ранг1;0));ЕСЛИ(ЕНД(ПОИСКПОЗ(H7;Ранг2;0));3;2);1);J5:L8)
16	-0,698709	0,156199	0,8934358	11	34	56			
17	-0,008464	-0,762591	3,0504438	29	10	90			
		1,433958	0,3616432	#ИД	69	42			
		1,064567	1,2220511	#ИД	58	64			
		0,106835	1,2869727	#ИД	31	65			
			1,9528609	#ИД	#ИД	85			
			0,7410673	#ИД	#ИД	51			

Рис. 12.1. Вычисление выборочной медианы

К сожалению, найти значение выборочной медианы не удастся без явного вычисления рангов значений объединенной выборки. Эти ранги вычисляются в столбцах

D:F с помощью формулы массивов $\{=РАНГ(A2:C41;A2:C41;1)\}$, охватывающей диапазон D2:F41. Здесь A2:C41 — диапазон ячеек, содержащий все выборочные значения. Поскольку выборки имеют разные объемы, часть ячеек диапазона D2:F41 будет содержать значение ошибки #Н/Д (в тех ячейках, которые соответствуют пустым ячейкам диапазона A2:C41), однако это не повлияет на последующие вычисления.

В ячейках H2:H4 вычисляются объемы выборок, а в ячейке H5 — объем объединенной выборки (формулы приведены на рис. 12.1). В ячейках H6 и H7 в зависимости от четности или нечетности значения объема объединенной выборки определяются ранги выборочных значений, по которым будет вычисляться выборочная медиана. Если значение объема объединенной выборки нечетно, то ранги будут совпадать.

Далее по этим значениям рангов надо найти соответствующие им выборочные значения. Для одной выборки сделать это несложно с помощью функции ВПР (подобные вычисления описаны в разделе 10.11). В данном случае применение функции ВПР затруднено (поскольку поиск необходимо вести не по одному столбцу, а по нескольким), но также возможно. Однако применим функцию БИЗВЛЕЧЬ (это функция из категории функций баз данных).

Синтаксис данной функции:

БИЗВЛЕЧЬ(База_данных;Поле;Критерий)

Эта функция в базе данных (диапазон ячеек, содержащий базу данных, задается первым аргументом функции) извлекает значение из указанного поля (второй аргумент) той записи, которая удовлетворяет критериям поиска (диапазон ячеек, содержащий критерий поиска, задается в качестве третьего аргумента). В данном случае сложность применения этой функции состоит в том, что заранее неизвестно, из какого поля (т.е. выборки) извлекать значение. Из этого положения можно выйти с помощью формулы (она записана в ячейке H8; формула в ячейке H9 практически совпадает с данной)

$$=БИЗВЛЕЧЬ(A1:F41;ЕСЛИ(ЕНД(ПОИСКПОЗ(H6;Ранги1;0));$$
$$ЕСЛИ(ЕНД(ПОИСКПОЗ(H6;Ранги2;0));3;2);1);J1:L4).$$

Здесь диапазоны ячеек, содержащие вычисленные ранги, названы соответственно Ранги1, Ранги2 и Ранги3.

Чтобы разобраться, как работает эта формула, разобьем ее на отдельные части. На рис. 12.2 такие части-формулы выполняются в ячейках H12:H16 (формула из ячейки H14 не используется в конечной формуле — она приведена для полноты картины). Формулы в ячейках H12:H14 определяют, какой выборке принадлежит выборочное значение с рангом, значение которого записано в ячейке H6. Они возвращают число, равное позиции выборочного значения в выборке, если этот ранг принадлежит данной выборке. В противном случае формула возвращает значение ошибки #Н/Д. (Для пояснения в соседних ячейках G12:G14 записаны числа (не формулы), соответствующие номеру выборки.) Таким образом, имеем “индикатор”, указывающий номер выборки, — выборочное значение с данным рангом принадлежит той выборке, для которой формула возвращает число, а не значение ошибки. На основании этого “индикатора” построена формула в ячейке H15, которая и вычисляет номер выборки (сравните значения в этой ячейке и в ячейках G12:G14). Здесь использована функция ЕНД, которая возвращает значение ИСТИНА, если ее аргумент имеет значение ошибки #Н/Д.

Наконец, формула в ячейке H16, аналогичная формуле в ячейке H8, возвращает выборочное значение, соответствующее указанному рангу. Номер выборки

	C	D	E	F	G	H		K	L	
1	Выборка3	Ранг1	Ранг2	Ранг3	Объемы			Ранг1	Ранг2	Ранг3
2	3,7450136	21	25	90	Выборка1	20		45		
3	0,8953411	22	19	53	Выборка2	30			45	
4	0,8836266	43	31	57	Выборка3	40				45
5	1,5926587	30	84	78	Всего	90	Ранг1	Ранг2	Ранг3	
6	1,5076078	47	5	73	Ранг1	45		46		
7	0,5278244	7	13	44	Ранг2	46			46	
8	0,2441714	51	27	36	Значение1	0,574582				46
9	1,5756723	80	61	78	Значение2	0,82354				
10	1,2050452	8	2	64	Медиана	0,589061				
11	0,9663213	6	38	56						
12	1,7790743	40		83	1	#И/Д	=ПОИСКПОЗ(\$H\$6;Ранг1;0)			
13	1,1222615	60	82	62	2	11	=ПОИСКПОЗ(\$H\$8;Ранг2;0)			
14	1,0102044	37	85	58	3	#И/Д	=ПОИСКПОЗ(\$H\$6;Ранг3;0)			
15	1,5005633	18	55	72		2	=ЕСЛИ(ЕНД(Н12);ЕСЛИ(ЕНД(Н13;G14;G13);G12			
16	-1,2691977	12	11	3	Значение1	0,574582	=БИЗВЛЕЧЬ(\$A\$1:\$F\$41;Н15;\$J\$1:\$L\$4)			
17	-0,7809815	70	15	50						

Итак, значение выборочной медианы подсчитано в ячейке НЮ. Очевидно, что вычисления в ячейках Н8 и Н9 являются промежуточными и от них можно освободиться, создав одну большую формулу для вычисления медианы. Однако такая формула будет практически не читаемой и станет источником потенциальных ошибок (хотя бы при вводе такой формулы). Чтобы немного освободить рабочий лист, столбцы D:F, содержащие ранги, можно скрыть, а ячейки с критериями для функции БИЗВЛЕЧЬ переместить "за экран" (на вычисления это не повлияет).

Глава 12. Сравнение одномерных выборок 353

На рис. 12.4 показан тот же рабочий лист для новых выборок, имеющих одинаковые стандартные нормальные распределения и нечетное число значений в объединенной выборке.

J4	{=СЧЕТ(ЕСЛИ(Выборка2>\$H\$10;Выборка2;"")}									
	G	H	I	J	K	L	M			
1	Объемы	Уровень значимости		0,05		=L3-J3				
2	Выборка1	20	Больше т. Менее т. Всего							
3	Выборка2	30	Выборка 1	8	12	20	=ЕСЛИ(И(ЕНЕЧЕТ(\$H\$5);\$H\$11=1);H2-1;H2)			
4	Выборка3	40	Выборка 2	8	22	30	=ЕСЛИ(И(ЕНЕЧЕТ(\$H\$5);\$H\$11=2);H3-1;H3)			
5	Всего	90	Выборка 3	29	11	40	=ЕСЛИ(И(ЕНЕЧЕТ(\$H\$5);\$H\$11=3);H4-1;H4)			
6	Ранг1	45	Всего	45	45	90	=СУММ(L3:L5)			
7	Ранг2	46	=СУММ(J3:J5)		{=СЧЕТ(ЕСЛИ(Выборка1>\$H\$10;Выборка1;"")}					
8	Значение1	0,280289	Критериальная статистика							
9	Значение2	0,275826	15,433333=((J3^2+K3^2)*2/L3+(J4^2+K4^2)*2/L4+(J5^2+K5^2)*2/L5)-L8							
10	Медиана	0,268057	Критические значения							
11	На выборки	3	т. нижнее	т. верхнее						
12			0,05063571	7,3777791						
13			=ХИ2ОБР(1-K/2,2)	=ХИ2ОБР(K/2,2)						
14			Гипотеза отклоняется							
15										
16			=ЕСЛИ(И(I12<J9,J9<J12);"принимается";"отклоняется")							
17			=ЕСЛИ(ЕНД(ПОИСКПОЗ(\$H\$6;Ранг1;0));ЕСЛИ(ЕНД(ПОИСКПОЗ(\$H\$6;Ранг2;0));3;2);1)							

Рис. 12.3. Рабочий лист для критерия медианы

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z											
1	Выборка1	Выборка2	Выборка3	Объемы		Уровень значимости		0,05																													
2	0,0486637	1,2667884	0,691113	Выборка1	20	Больше т		Меньше т		Всего																											
3	-0,519457	1,5053077	0,1235652	Выборка2	30	Выборка 1	8	12	20																												
4	-0,218865	0,9233611	0,2859425	Выборка3	41	Выборка 2	18	11	29																												
5	2,5375406	1,7756083	1,6995294	Всего	91	Выборка 3	19	22	41																												
6	-0,271452	-2,482381	-1,295105	Ранг1	48	Всего	45	45	90																												
7	-1,423009	-0,537718	-0,834172	Ранг2	46																																
8	-1,968034	-0,85412	1,7052164	Значение1	-0,07388	Критериальная статистика																															
9	-0,074908	-0,752612	-0,437052	Значение2	-0,07388	2,7091874																															
10	-0,01826	-0,843736	-0,480299	Медиана	-0,07388	Критические значения																															
11	-0,943346	0,260398	1,8174623	На выборки	2	т. нижнее		т. верхнее																													
12	0,8711673	0,0994799	0,4363539																																		
13	-0,72043	0,9680893	0,2850222																																		
14	-0,426843	-1,743853	-0,015891																																		
15	0,7047779	0,7824717	-0,114801																																		
16	0,0365731	-0,352287	1,303239																																		

Рис. 12.4. Критерий медианы для новых данных

Реализация критерия для дискретных распределений

Здесь также сначала подсчитаем значение выборочной медианы. На рис. 12.5 показаны две выборки объемом соответственно 30 и 50 значений, представленных в виде частотных таблиц и имеющих распределение Пуассона с параметром $X = 2$. Выборки получены с помощью средства Генерация случайных чисел, затем для них подсчитаны частотные таблицы так, как показано в разделе 8.3.1. По этим частотным таблицам подсчитываются частоты и накопленные частоты объединенной выборки: частоты просто складываются для одинаковых значений, а накопленные частоты вычисляются так, как описано в разделе 8.3.1. Способ

вычисления медианы описан выше, в разделе Вычисления в п. 1, Б. Формулы, необходимые для вычисления медианы, показаны на рис. 12.5. Здесь интервал $G_3:G_{10}$, содержащий значения накопленных частот, назван Нч. Также отметим формулы в ячейках 13 и 14, в данном случае выполняющие роль функции ВПР, которую нельзя применить непосредственно без перестановки столбцов так, чтобы столбец с накопленными частотами предшествовал столбцу со значениями.

G4		=F4+G3							
A		B		C		D		E	
Выборка 1		Выборка 2		Объединенная выборка		F(m)		35	
Значения	Частоты	Значения	Частоты	Значения	Частоты	Накопленные частоты	F(m+1)	x(m)	
0	6	0	8	0	14	14		1	
1	8	1	13	1	21	35		2	
2	4	2	12	2	16	51		Медиана	1,09804
3	4	3	8	3	12	63			
4	6	4	5	4	11	74			
5	0	5	4	5	4	78			
6	1			6	1	79			
7	1			7	1	80			
Выборка1		30		Выборка2		50		Всего	
				Формулы в ячейке 11		={МАКС(ЕСЛИ(НЧ<G11/2;НЧ;0))}			
				в ячейке 12		={МИН(ЕСЛИ(НЧ>G11/2;НЧ;1000))}			
				в ячейке 13		={СУММ(ЕСЛИ(НЧ=1;ЕЗ;Е10;""))}			
				в ячейке 14		={СУММ(ЕСЛИ(НЧ=2;ЕЗ;Е10;""))}			
				в ячейке 15		={3+(14-13)*(G11/2-1)/2}			

Рис. 12.5. Вычисление медианы для дискретных распределений

На рис. 12.6 показан законченный рабочий лист и формулы для вычислений. Отметим, что суммы значений по столбцам в таблице K3:L5, в отличие от случая непрерывных распределений, не обязаны равняться $[n/2]$, поскольку здесь выборочная медиана вычисляется без привлечения порядковых статистик. По этой же причине значение выборочной медианы только в исключительных случаях (когда накопленная частота для первого значения равна $na/2$) будет совпадать с выборочными значениями. Данное обстоятельство значительно облегчает вычисление R_s . Если значение выборочной медианы все-таки совпадает с каким-либо выборочным значением, можно немного увеличить значение выборочной медианы (это не повлияет на результат вычисления критерия).

12.1.2. Критерий Уилкоксона—Манна—Уитни

Этот критерий является модификацией критерия знаковых рангов Уилкоксона для проверки гипотез о значении медианы (см. раздел 11.4.2), обобщенного для случая двух выборок¹. Кратко критерий описан в разделе 2.4.2.

¹ Этот критерий и его модификации также называют критерием Уилкоксона и критерием Манна-Уитни. Первоначально он был разработан Уилкоксоном (Wilcoxon, 1945 г.) для выборок одинаковых объемов, а затем обобщен для случая выборок произвольных объемов Манном и Уитни (Mann, Whitney, 1947 г.).

K5		=СУММ(K3:K4)	
1	F(m)	35	Уровень значимости 0,05
2	F(m+1)	51	
3	x(m)	1	Выборка 1
4	x(m+1)	2	Выборка 2
5	Медиана	1,098	Всего
6	=СУММ(ЕСЛИ(Значения2<4;Частоты2;""))		
7	Критериальная статистика		
8	1,413333		
9	Критические значения		
10	t нижнее t верхнее		
11	0,0009821 5,023903		
12	=ХИ2ОБР(1-L/2;1)		
13	Гипотеза принимается		
14	=ЕСЛИ(И(J11<K8;K8<K11),"принимается","отклоняется")		

Рис. 12.6. Рабочий лист критерия медианы для сравнения дискретных распределений

Статистическая модель. Даны две одномерные независимые выборки объемом соответственно n_1 и n_2 , имеющие непрерывные распределения.

Гипотезы

H_0 : выборки имеют одинаковые распределения;

H_1 : нулевая гипотеза неверна.

Задан уровень значимости α .

Вычисления

1. Обе выборки объединяются в единую выборку, и по объединенной выборке строится вариационный ряд $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ и вычисляются ранги выборочных значений. Если встречаются одинаковые значения, то им приписываются равные средние ранги. Здесь и далее $n = n_1 + n_2$.
2. Для одной из выборок подсчитывается сумма рангов R , которые получили ее выборочные значения в объединенной выборке. Если обозначить через R_1 сумму рангов первой выборки, а через R_2 — сумму рангов второй выборки, то эти суммы будут связаны соотношением $R_1 + R_2 = n(n+1)/2$. Поэтому достаточно вычислить сумму рангов только одной выборки. Обычно вычисляется сумма рангов выборки, имеющей меньший объем, а сумма рангов другой выборки вычисляется на основании приведенного соотношения.
3. Вычисляется критериальная статистика.

а) Для малых выборок:

$$U_1 = n_1 n_2 + \frac{1}{2} n_1 (n_1 + 1) - R_1, \quad U_2 = n_1 n_2 + \frac{1}{2} n_2 (n_2 + 1) - R_2, \quad U = \max(U_1, U_2).$$

б) Для больших выборок:

$$T = \frac{U - \frac{1}{2} n_1 n_2}{\sqrt{\frac{n_1 n_2 (n+1)}{12}}}.$$

Построение критической области. При условии истинности нулевой гипотезы статистика U имеет специальное распределение Манна–Уитни, а статистика T имеет асимптотически стандартное нормальное распределение.

- а) Для малых выборок вычисляется критическое значение t как квантиль порядка $1 - \alpha/2$ распределения Манна–Уитни. Гипотеза H_0 принимается, если выполняется неравенство $U \leq t$, иначе гипотеза H_0 отклоняется.
- б) Для больших выборок вычисляется критическое значение t как квантиль порядка $1 - \alpha/2$ стандартного нормального распределения. Гипотеза H_0 принимается, если выполняется неравенство $|T| \leq t$, иначе гипотеза H_0 отклоняется.

Комментарии

1. Для малых выборок, несмотря на то что вычисляется только одно критическое значение, критерий является двухсторонним с уровнем значимости α .
2. Критерий считается более точным, чем критерий медианы.
3. Существуют различные мнения о том, какого объема выборки достаточно для того, чтобы применять нормальную аппроксимацию. “Средняя” оценка — объем каждой выборки должен быть не менее 20.
4. Таблицы со значениями квантилей распределения Манна–Уитни приводятся во многих источниках, например [4, 9, 14]. В [18] показан способ вычисления этих квантилей.

Практическая реализация

Покажем реализацию критерия с использованием критериальной статистики T , т.е. с использованием нормальной аппроксимации. Все формулы, необходимые для вычислений, показаны на рис. 12.7. В качестве тестовых взяты две выборки, имеющие нормальное распределение: одна — стандартное, вторая — с единичными математическим ожиданием и дисперсией. Диапазоны ячеек, содержащие выборочные значения, названы соответственно Выборка1 и Выборка2. В столбцах C и D вычислены ранги значений объединенной выборки с помощью формулы массива $\{=РАНГ(A2:B41;A2:B41;1)\}$ (в диапазоне A2:B41 содержатся значения обеих выборок). Диапазоны ячеек, содержащие ранги для первой и второй выборок, названы соответственно Ранг1 и Ранг2. Как видно на рис. 12.7, нулевая гипотеза о совпадении распределений в данном случае отвергается.

12.1.3. Критерий Краскала–Уоллиса

Этот критерий является обобщением критерия Уилкоксона–Манна–Уитни для случая нескольких (более двух) выборок.

Статистическая модель. Даны k одномерных независимых выборок объемом соответственно n_1, n_2, \dots, n_k , имеющих непрерывные распределения.

Гипотезы

H_0 : все k выборок имеют одинаковые распределения;

H_1 : нулевая гипотеза неверна.

Задан уровень значимости α .

C2		{=РАНГ(A2:B41;A2:B41;1)}								
	A	B	C	D	E	F	G	H	I	J
1	Выборка	Выборка	Ранг1	Ранг2	Объемы			Уровень значимости		
2	1.54629	2.23478	51	68	Выборка1	30	=СЧЕТ(Выборка1)		0.05	
3	0.28295	-0.4426	29	13	Выборка2	40	=СЧЕТ(Выборка2)	Критериальная статистика		
4	1.14842	0.28944	57	30	Всего	70	=СУММ(F2:F3)		2.61092	
5	-0.304	-1.2044	17	4	R1	1285	=СУММ(Ранг1)	Критическое значение		
6	-0.1478	2.05751	19	67	R2	1200	=СУММ(Ранг2)		1.95996	
7	-0.3808	0.18147	16	27	U1	380		Гипотеза отклоняется		
8	2.75234	0.92833	70	46	U2	820	=F2*F3-F5+F2*(F2+1)/2			
9	1.74416	0.37708	63	34	U	820	=F2*F3-F6+F3*(F3+1)/2			
10	0.1196	0.98272	25	48			=МАКС(F7,F8)			
11	1.09151	0.37489	55	33						
12	1.85606	1.03487	62	52	В ячейке I4	=(F9-F2*F3/2)/КОРЕНЬ(F2*F3*(F4+1)/12)				
13	-0.4462	-0.0532	12	23	В ячейке I6	=НОРМСТОБР(1-I2/2)				
14	0.98547	-0.6986	49	8	В ячейке I7	=ЕСЛИ(ABS(I4)<I6,"принимается","отклоняется")				
15	1.78871	0.9301	64	47						
16	0.55101	-1.0297	36	6						

Рис. 12.7. Критерий Уилкоксона-Манна-Уитни

Вычисления

1. Все k выборок объединяются в единую выборку, и по объединенной выборке строится вариационный ряд $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ и находятся ранги выборочных значений. Если встречаются одинаковые значения, то им приписываются равные средние ранги. Здесь и далее $n = \sum_{i=1}^k n_i$.
2. Для каждой из выборок подсчитывается сумма рангов R_i , которые получили ее выборочные значения в объединенной выборке.
3. Вычисляется критериальная статистика $T = \frac{12}{n(n+1)} \sum_{i=1}^k \frac{R_i^2}{n_i} - 3(n+1)$.

Построение критической области. При условии истинности нулевой гипотезы статистика T имеет специальное распределение Краскала-Уоллиса. Асимптотически эта статистика T имеет распределение χ^2 с $(k-1)$ степенью свободы.

- а) Для малых выборок вычисляется критическое значение t как квантиль порядка $1 - \alpha$ распределения Краскала-Уоллиса. Гипотеза H_0 принимается, если выполняется неравенство $T \leq t$, иначе гипотеза H_0 отклоняется.
- б) Для больших выборок вычисляется критическое значение t как квантиль порядка $1 - \alpha$ распределения χ^2 с $(k-1)$ степенью свободы. Гипотеза H_0 принимается, если выполняется неравенство $T \leq t$, иначе гипотеза H_0 отклоняется.

Комментарии

1. Таблицы со значениями квантилей распределения Краскала-Уоллиса приводятся в [9, 14].
2. В [22] показана более точная аппроксимация, основанная на преобразовании статистики T с использованием F -распределения.

3. Критерий считается более точным, чем критерий медианы.

4. Если нулевая гипотеза отклоняется, то критерий не позволяет определить, какие совокупности имеют различные распределения. Однако применение для определения различных распределений попарных сравнений выборок методом Уилкоксона–Манна–Уитни нежелательно, поскольку при многократном применении одного критерия резко возрастает вероятность ошибки первого рода.

Практическая реализация этого критерия почти полностью совпадает с реализацией метода Уилкоксона–Манна–Уитни (за исключением вычисления критериальной статистики и критического значения).

12.1.4. Критерий серий Вальда–Вольфовица

Статистическая модель. Даны две одномерные независимые выборки объемом соответственно n_1 и n_2 , имеющие непрерывные распределения.

Гипотезы

H_0 : выборки имеют одинаковые распределения;

H_1 : нулевая гипотеза неверна.

Задан уровень значимости α .

Вычисления

1. Обе выборки объединяются в единую выборку, и по объединенной выборке строится вариационный ряд $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ ($n = n_1 + n_2$).

2. По вариационному ряду подсчитывается количество *серий* U — количество участков вариационного ряда, в которых присутствуют значения только одной выборки.

3. Вычисляется критериальная статистика $T = \frac{U - \frac{2n_1n_2}{n} + 1}{\sqrt{\frac{2n_1n_2(2n_1n_2 - n)}{n^2(n+1)}}}$.

Построение критической области. При условии истинности нулевой гипотезы статистика T имеет асимптотически стандартное нормальное распределение. Вычисляется критическое значение t как квантиль порядка $1 - \alpha/2$ стандартного нормального распределения. Гипотеза H_0 принимается, если выполняется неравенство $|T| \leq t$, иначе гипотеза H_0 отклоняется.

Комментарии

1. Этот критерий не так чувствителен к форме распределений, как к параметрам положения распределений. Поэтому его часто используют как непараметрический критерий сравнения математических ожиданий двух выборок.

2. Критерий является асимптотическим. Он применяется, если каждая из выборок имеет больше 20 значений.

Практическая реализация

На рис. 12.8 показан рабочий лист, реализующий критерий серий. В столбцах А и В содержатся две выборки объемом соответственно 20 и 30 значений. В столбцах С и D, как и в критерии Уилкоксона–Манна–Уитни, подсчитаны ранги

значений объединенной выборки. Затем эти ранги отсортированы по возрастанию, причем каждый столбец в отдельности (можно сортировать по убыванию — это не существенно). Значения рангов, идущие в натуральном порядке, образуют серии. Эти серии на рис. 12.8 показаны разными цветами. Теперь надо подсчитать количество серий. Для этого в соседних столбцах в ячейке E2 и F2 введены значения 1. Далее в ячейках E3 и F3 введены формулы =ЕСЛИ(C3=C2+1;0;1) и =EGTM(D3=D2+1;0;1) соответственно, которые затем скопированы вниз. Таким образом ставится единица в начале серии, а остальным элементам серии ставится в соответствие нуль. Количество серий подсчитывается в ячейке H5 как сумма единиц в диапазонах Серии1 и Серии2. Остальные формулы критерия, вычисляющие критериальную статистику и критическое значение, показаны на рис. 12.8.

E3 =ЕСЛИ(C3=C2+1;0;1)

	A	B	C	D	E	F	G	H	I	J	K
1	Выборка	Выборка	Ранг1	Ранг2	Серии1	Серии2	Объемы	Уровень значимости			
2	-0,04679	-0,76739	4	1	1	1	Выборка1	20		0,1	
3	2,071159	-0,7281	16	1	1	0	Выборка2	30			
4	2,017747	-0,14225	20	1	1	0	Всего	50			
5	0,35671	-1,23473	21	0	0	1	К-во серий	28	=СУММ(Серии1;Серии2)		
6	1,626953	0,124722	23	1	1	0	Статистика	0,91			
7	2,329809	0,327302	24	0	0	0	Критическое значение				
8	1,057509	-0,88678	27	1	1	0		1,64			
9	1,622072	1,777066	28	0	0	0	Гипотеза	принимается			
10	4,100787	0,377388	33	1	1	0					
11	2,362824	-1,51969	34	1	1	0					
12	0,428261	2,249709	39	1	1	0					
13	0,852859	-0,86938	40	1	1	0					
14	2,267938	1,551569	41	0	0	0					
15	2,339863	-0,3254	43	1	1	0					
16	-0,64028	-0,00315	50	0	0	1	"отклоняется"				
17	1,585178	-0,26843	50	1	1	0					
18	0,005186	-0,30356	56	1	1	0					
19	1,178443	-0,58064	56	1	1	1					
20	1,561867	0,785915		1	1	1					

$$=(H5-1-2 \cdot H2 \cdot H3 / (H4)) \cdot H4 /$$

$$\text{КОРЕНЬ}(2 \cdot H2 \cdot H3 \cdot (2 \cdot H2 \cdot H3 - H4) / (H4 + 1))$$

$$=ЕСЛИ(ABS(H6) < H8, "принимается",$$

Рис. 12.8. Критерий серий

12.1.5. Критерий %²

Данный критерий является обобщением для случая нескольких выборок одноименного критерия, описанного в разделах 2.4.3 и 9.3. Критерий можно применять для сравнения как непрерывных, так и дискретных распределений. Однако чаще его применяют для сравнения дискретных распределений. В случае непрерывных распределений, если определены интервалы, на которые разбивается область возможных выборочных значений, и подсчитаны частоты попадания выборочных значений в эти интервалы (см. раздел 9.3), критериальные вычисления совпадают с аналогичными вычислениями для дискретных распределений. Поэтому опишем данный критерий для случая дискретных распределений.

Статистическая модель. Даны k одномерных независимых выборок объемом соответственно n_1, n_2, \dots, n_k , имеющих дискретные распределения. Предполагается, что выборки заданы в виде частотных таблиц. (О вычислении частотных таблиц речь идет в разделе 8.3.1.)

Гипотезы

H_0 : все h выборок имеют одинаковые распределения;

H^{\wedge} нулевая гипотеза неверна.

Задан уровень значимости α .

Вычисления

1. Все k частотных таблиц объединяются в единую таблицу следующего вида.

Значения	X/	x_2	...	x_n	Всего
Частоты выборки 1	f_n	f_{12}	...	f_{1m}	*
					$\tau=1$
Частоты выборки 2	f_{21}	f_{22}	-	f_{2m}	Л
					$\tau=1$
Частоты выборки k	f_{k1}	f_{k2}	...	f_{km}	ТМ
					$\tau=1$
Всего	*	*		*	*
	$i=1$	$i=1$		$i=1$	$i=1$

2. Вычисляется критериальная статистика $T = n \left(\sum_{j=1}^m \sum_{i=1}^k \frac{f_{ij}^2}{n_i f_j} - 1 \right)$.

Построение критической области. При условии истинности нулевой гипотезы статистика T асимптотически имеет распределение χ^2 с $(m-1)(k-1)$ степенью свободы.

Вычисляются критические значения t_{kr} как квантиль порядка $1-\alpha$ распределения χ^2 с $(m-1)(k-1)$ степенью свободы. Гипотеза H_0 принимается, если выполняется неравенство $T \leq t_{kr}$, иначе гипотеза H_0 отклоняется.

Комментарий. Критерий является асимптотическим. Считается, что его можно применять, если объем каждой выборки — не менее 15.

Практическая реализация

На рис. 12.9 показан рабочий лист с двумя выборками, представленными в виде частотных таблиц, которые использовались в примере критерия медианы для дискретных распределений. В столбце E вычисляются суммы частот для каждого значения (соответствуют строке Всего в приведенной выше таблице). В ячейке G4 вычисляется критериальная статистика с помощью формулы массива, показанной на рис. 12.9. В случае большого количества выборок эту формулу, по-видимому, следует разбить на несколько (по количеству выборок), чтобы сделать ее более читаемой и простой.

	E3		f _в = B3+D3					
	A	B	C	D	E	F	G	H
1	Выборка 1		Выборка 2		Уровень значимости			
2	Значения	Частоты	Значения	Частоты	n=	0,05		
3	0	6	0	8	14	Критериальная статистика		
4	1	8	1	13	21	8,427128		
5	2	4	2	12	16	Критическое значение		
6	3	4	3	8	12	14,06713 =ХИ2ОБР(G2;7)		
7	4	6	4	5	11	Гипотеза принимается		
8	5	0	5	4	4	=ЕСЛИ(G4<G6;"принимается";"отклоняет")		
9	6	1			1			
10	7	1			1			
11	n1= 30		n2= 50		80			
12								
13	=E11*(СУММПРОИЗВ((B3:B10)^2,1/E3:E10)/B11+СУММПРОИЗВ((D3:D10)^2,1/E3:E10)/D11-1))							
14								
15								

Рис. 12.9. Критерий χ^2

12.1.6. Критерий Смирнова

Этот критерий, как и критерий Колмогорова (см. раздел 9.3), построен на сравнении не отдельных параметров распределения (приведенные выше критерии основаны на сравнении местоположения распределений), а функций распределения. В данном критерии сравниваются эмпирические функции распределения.

Статистическая модель. Даны две одномерные независимые выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_m объемом соответственно n и m , имеющие непрерывные распределения.

Гипотезы

H_0 : выборки имеют одинаковые распределения;

H_1 : нулевая гипотеза неверна.

Задан уровень значимости α .

Вычисления

1. По каждой выборке в отдельности строятся вариационные ряды $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ и $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(m)}$.

2. Вычисляются разности $D_{m,r}^+ = \frac{r}{m} - F_x(y_{(r)})$ (или $D_{n,s}^+ = F_y(x_{(s)}) - \frac{s-1}{n}$)

и $D_{m,r}^- = F_x(y_{(r)}) - \frac{r-1}{m}$ (или $D_{n,s}^- = \frac{s}{n} - F_y(x_{(s)})$), $r = 1, 2, \dots, n$, $s = 1, 2, \dots, m$. Здесь

F_x и F_y — эмпирические функции распределения соответственно первой и второй выборок. Можно вычислять разности $D_{m,r}^+$ и $D_{m,r}^-$ либо разности $D_{n,s}^+$ и $D_{n,s}^-$.

3. Вычисляется критериальная статистика по формуле $D_{n,m} = \max_{1 \leq r \leq m} (D_{m,r}^+, D_{m,r}^-)$ либо $D_{n,m} = \max_{1 \leq s \leq n} (D_{n,s}^+, D_{n,s}^-)$.

Построение критической области. При условии истинности гипотезы H_0 статистика $D_{n,m}$ имеет так называемое распределение Смирнова.

Вычисляется критическое значение $t_{кр}$ — квантиль порядка $1 - \alpha$ распределения Смирнова. Гипотеза H_0 принимается, если $D_{n,m} \leq t_{кр}$. В противном случае гипотеза H_0 отклоняется.

Комментарии

1. Для вычисления квантилей распределения Смирнова существуют специальные таблицы, которые приведены во многих книгах по математической

статистике. Поскольку случайная величина $\sqrt{\frac{nm}{n+m}} D_{n,m}$ асимптотически имеет распределение Колмогорова, при $n, m \geq 40$ и $0,01 \leq \alpha \leq 0,2$ можно воспользоваться приближенной формулой для вычисления $t_{кр}$:

$$t_{кр} = \sqrt{\frac{-\ln(0,5\alpha)}{2N}} - \frac{1}{6N}, \text{ где } N = \frac{nm}{n+m} \quad [4].$$

2. Существует много преобразований величины $D_{n,m}$ и аппроксимаций распределения Смирнова, которые позволяют не обращаться непосредственно к распределению Смирнова для нахождения критических значений. Подробное описание критерия Смирнова и его вариантов можно найти в [22].

Практическая реализация

На рис. 12.10 показан рабочий лист Excel, реализующий критерий Смирнова. В качестве тестовых выборок взяты две выборки, имеющие стандартное нормальное распределение. Критериальные вычисления значительно упрощаются, если выборки предварительно отсортированы по возрастанию, как и сделано на предлагаемом рабочем листе. В столбцах C и D подсчитаны ранги выборочных значений. Для этого можно использовать функцию РАНГ или, предполагая, что нет совпадающих значений, можно просто ввести последовательность натуральных чисел.

	A	B	C	D	E	F	G	H	I	J	K
1	Выборка1	Выборка2	Ранг1	Ранг2	Дминус	Дплюс	Объемы	Уровень значимости			
2	-2,6547	-1,86634	1	1	0,025	0	Выборка1	40	0,05		
3	-1,7971	-1,47901	2	2	0,0167	0,0083	Выборка2	30	=СЧЕТ(Выборка2)		
4	-1,6239	-1,14592	3	3	0,0417	-0,017	D=	0,283	=МАКС(Дплюс;Дминус)		
5	-1,2891	-0,61049	4	4	0,0333	-0,008	N=	17,14	=H2*H3/(H2+H3)		
6	-1,04	-0,56191	5	5	0,025	0	Критическое значение				
7	-0,8894	-0,33027	6	6	0,05	-0,025		0,318	=КОРЕНЬ(-LN(0,5*2)/(2*H5))-1/(6*H5)		
8	-0,8611	-0,15086	7	7	0,075	-0,05	Гипотеза	принимается			
9	-0,7561	-0,14113	8	8	0,1	-0,075					
10	-0,5761	-0,0146	9	9	0,0917	-0,067	=ЕСЛИ(H4<H7,"принимается","отклоняется")				
11	-0,3719	-0,00513	10	10	0,0833	-0,058					
12	-0,3706	0,01356	11	11	0,1083	-0,083	=ЕСЛИ(A12<=МИН(Выборка2);0;ВПР(A12;\$B\$2:\$D\$31;3;1)/\$H\$3)				
13	-0,3607	0,03825	12	12	0,1333	-0,108					
14	-0,329	0,12977	13	13	0,125	-0,1	=C14/\$H\$2-ЕСЛИ(A14<=МИН(Выборка2);0;ВПР(A14;\$B\$2:\$D\$31;3;1)/\$H\$3)				
15	-0,3122	0,27093	14	14	0,15	-0,125					
16	-0,2744	0,29991	15	15	0,175	-0,15					
17	-0,2649	0,31775	16	16	0,2	-0,175					

Рис. 12.10. Критерий Смирнова

В столбцах E и F вычисляются значения разностей $D_{n,s}^-$ и $D_{n,s}^+$. Опишем, как вычисляются разности $D_{n,s}^-$. Сначала в ячейку E2 вводится формула

$$=C2/\$H\$2-ЕСЛИ(A2<=МИН(Выборка2);0;ВПР(A2;\$B\$2:\$D\$31;3;1)/\$H\$3).$$

1. Построенный доверительный интервал является приближенным. Если нет оснований отвергать предположение о равенстве дисперсий, то предпочтительнее использовать точный доверительный интервал из предыдущего раздела.
2. Если известны значения дисперсий σ_x^2 и σ_y^2 , то вместо распределения Стьюдента используется стандартное нормальное распределение, а в формуле вычисления $A_{n,m}$, S_x^2 и S_y^2 заменяются значениями σ_x^2 и σ_y^2 .
3. Описанный метод построения доверительного интервала устойчив при умеренных отклонениях от нормальности.
4. Для достаточно больших объемов выборок, например при $n + m > 30$, вместо распределения Стьюдента можно использовать стандартное нормальное распределение.

Практическая реализация

На рис. 12.11 показан рабочий лист Excel, реализующий данный метод построения доверительного интервала для разности математических ожиданий. Все формулы, по которым выполняются вычисления, показаны на этом рисунке. Отметим, что первая выборка имеет стандартное нормальное распределение, а вторая — нормальное распределение с единичным математическим ожиданием и дисперсией, равной 4. Таким образом, здесь $\delta = -1$. Также обращаем внимание на первый аргумент функции СТЬЮДРАСПОБР — эта функция не является обратной к функции распределения Стьюдента, а находит корень уравнения $P(X \geq u) = p$ (см. раздел 4.7.7).

E2		=СЧЕТ(Выборка1)							
A	B	C	D	F	G	H	I		
Выборка1	Выборка2	Объемы выборок		Доверительный уровень					
-0,545376	-1,990282	Выборка1	20		0,95				
-0,571722	1,1430857	Выборка2	30	=СЧЕТ(Выборка2)					
1,875599	2,5370323	Среднее1	-0,265730856	=СРЗНАЧ(Выборка1)					
-1,316904	3,39272	Среднее2	1,233412146	=СРЗНАЧ(Выборка2)					
0,681731	-1,035192	Дисперсия1	0,775778093	=ДИСПР(Выборка1)					
0,018384	3,7830167	Дисперсия2	4,072820469	=ДИСПР(Выборка2)					
0,637966	1,1936253	Дельта	-1,499143002	=E4-E5					
0,216684	0,6639739	Ап,м	0,41779132	=КОРЕНЬ(E6/E2+E7/E3)					
-1,372752	1,0901734	N	40,62754103	=-2+((E6/E2+E7/E3)^2)/((E6/E2)^2*(E2-1)+((E7/E3)^2*(E3-1)))					
-0,093668	-0,035397	k	2,328933988	=((E7/E3)^2*(E3-1))					
-0,099368	0,677536	Доверительный интервал		=СТЮДРАСПОБР((1-G2)/2,E10)					
-1,961804	2,2173756		-2,47215141	-0,526134597	=E8+E11*E9				
0,471934	-1,258228		* =E8-E11*E9						
-0,386432	-0,085047								
0,051855	3,8458313								

Рис. 12.11. Построение доверительного интервала для разности математических ожиданий двух выборок

12.2.3. Доверительный интервал для отношения дисперсий нормальных совокупностей

Статистическая модель. Даны две одномерные независимые выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_m объемом соответственно n и m , имеющие нормальное распределение.

Доверительный интервал для отношения $\gamma = \sigma_x^2 / \sigma_y^2$ дисперсий σ_x^2 и σ_y^2 соответственно первой и второй выборки строится следующим образом.

1. Вычисляются точечные оценки дисперсий $S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$

и $S_y^2 = \frac{1}{m} \sum_{i=1}^m (y_i - \bar{y})^2$ и дополнительно $B_{n,m} = \frac{(m-1)nS_x^2}{(n-1)mS_y^2}$.

2. Задается доверительный уровень α .

3. Вычисляются квантили t_1 и t_2 порядка соответственно $(1 - \alpha)/2$ и $(1 + \alpha)/2$ F-распределения со степенями свободы $n - 1$ и $m - 1$.

4. Вычисляется доверительный интервал: $\left(\frac{B_{n,m}}{t_2}, \frac{B_{n,m}}{t_1} \right)$.

Комментарий. Описанный метод построения доверительного интервала не устойчив при отклонениях от нормальности.

Практическая реализация

На рис. 12.12 показан рабочий лист Excel, реализующий данный метод построения доверительного интервала для отношения дисперсий. Все формулы, по которым выполняются вычисления, показаны на этом рисунке. Отметим, что первая выборка имеет стандартное нормальное распределение, а вторая — нормальное распределение с единичным математическим ожиданием и дисперсией, равной 4. Таким образом, здесь $\gamma = 0,25$.

E2		=СЧЕТ(Выборка1)			
1	Выборка1	Выборка2	Объемы выборки	Доверительный уровень	
2	0,292641	1,7068008	Выборка1	20	0,95
3	-0,477492	0,4319272	Выборка2	30	=СЧЕТ(Выборка2)
4	0,781416	0,4238175	Дисперсия1	0,948999963	=ДИСПР(Выборка1)
5	0,465505	1,0074912	Дисперсия2	4,901888327	=ДИСПР(Выборка2)
6	-0,748391	1,1316871	Bn,m	0,196580161	=(E3-1)*E2*E4/((E2-1)*E3*E5)
7	0,894149	1,7435866	t1	2,231274721	=ФРАСПОБР((1-G2)/2;E2-1;E3-1)
8	0,475779	-1,367458	t2	0,416330082	=ФРАСПОБР((1+G2)/2;E2-1;E3-1)
9	1,080965	-0,673292	Доверительный интервал		
10	-0,379746	3,2554031	0,08810218	0,472173811	=E6/E8
11	0,660264	1,4400521	* =E6/E7		
12	-0,93346	2,718629			

Рис. 12.12. Построение доверительного интервала для отношения дисперсий двух выборок

12.2.4. Доверительный интервал для разности двух биномиальных вероятностей

Статистическая модель. Имеются две серии наблюдений за экспериментом. В первой серии в каждом эксперименте с вероятностью p_1 происходит событие "1" ("успех"), во второй серии это событие происходит с вероятностью p_2 . Пусть

в первой серии зафиксировано n экспериментов, из них в r_1 случаях наблюдалось событие "1". Во второй серии зафиксировано m экспериментов, из них в r_2 случаях наблюдалось событие "1". Размеры серий больше 20.

Доверительный интервал для разности $\delta = p_1 - p_2$ строится следующим образом.

1. Вычисляются точечные оценки вероятностей p_1 и p_2 : $\hat{p}_1 = r_1/n$, $\hat{p}_2 = r_2/m$

$$\text{и дополнительно } \hat{\delta} = \hat{p}_1 - \hat{p}_2, A_{n,m} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}.$$

2. Задается доверительный уровень α .
3. Вычисляется значение k как квантиль порядка $(1 + \alpha)/2$ стандартного нормального распределения.
4. Вычисляется доверительный интервал: $(\hat{\delta} - kA_{n,m}, \hat{\delta} + kA_{n,m})$.

Комментарий. Описанный метод построения доверительного интервала является приближенным и основывается на аппроксимации биномиального распределения нормальным. Отсюда требование, чтобы объемы выборок были не меньше 20.

Практическая реализация в Excel данного метода не вызывает затруднений.

12.3. Проверка гипотез о параметрах распределений

В данном разделе сравниваются выборки в виде критериев проверки гипотез о равенстве или различии параметров распределений отдельных одномерных выборок. В этом случае, как и при построении доверительных интервалов для разностей или отношений параметров распределений, большую роль играют предположения о типе выборочных распределений. Большинство критериев, описанных ниже, относится к нормальным совокупностям и проверяет совпадение или математических ожиданий, или дисперсий.

12.3.1. Проверка гипотез о математических ожиданиях нормальных распределений

В этом разделе будут приведены критерии для сравнения математических ожиданий двух или более выборок, имеющих нормальное распределение.

Критерий проверки гипотезы о равенстве математических ожиданий при известных дисперсиях

Этот критерий описан в разделе 2.4.2 и в разделе 5.6, посвященном средству Двухвыборочный z-тест для средних.

Статистическая модель. Выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_m объемом соответственно n и m извлечены из совокупностей, имеющих нормальные распределения с известными дисперсиями σ_1^2 и σ_2^2 и математическими ожиданиями μ_1 и μ_2 соответственно.

Гипотезы

а) Равенство

б) Неравенство

$$H_0: \mu_1 = \mu_2$$

$$H_0: \mu_1 \leq \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

$$H_1: \mu_1 > \mu_2$$

Задан уровень значимости α .

Вычисления

1. По каждой выборке вычисляются выборочные средние $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$,

$$\bar{y} = \frac{1}{m} \sum_{i=1}^m y_i.$$

2. Вычисляется критериальная статистика $z = \frac{(\bar{x} - \bar{y})}{\sqrt{\sigma_1^2/n + \sigma_2^2/m}}$.

Построение критической области. При условии истинности нулевых гипотез статистика z имеет стандартное нормальное распределение.

Случай а). Вычисляется критическое значение $z_{кр}$ как квантиль порядка $1 - \alpha/2$ стандартного нормального распределения. Гипотеза H_0 принимается, если выполняется неравенство $|z| \leq z_{кр}$, иначе гипотеза H_0 отклоняется.

Случай б). Вычисляется критическое значение $z_{кр}$ как квантиль порядка $1 - \alpha$ стандартного нормального распределения. Гипотеза H_0 принимается, если $T \leq t$.

Комментарий. Критерий устойчив при умеренных отклонениях распределения выборки от нормального.

Практическая реализация в Excel этого критерия осуществляется с помощью средства Двухвыборочный z-тест для средних, которое описано в разделе 5.6. Там же приведен пример реализации критерия.

Критерий Стьюдента проверки гипотезы о равенстве математических ожиданий (случай равных дисперсий)

Этот критерий описан в разделе 2.4.2 и в разделе 5.7, посвященном средству пакета анализа Двухвыборочный t-тест с одинаковыми дисперсиями.

Статистическая модель. Выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_m объемом соответственно n и m извлечены из совокупностей, имеющих нормальные распределения с неизвестными, но равными дисперсиями σ^2 и математическими ожиданиями соответственно μ_1 и μ_2 .

Гипотезы

а) Равенство

б) Неравенство

$$H_0: \mu_1 = \mu_2$$

$$H_0: \mu_1 \leq \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

$$H_1: \mu_1 > \mu_2$$

Задан уровень значимости α .

Вычисления

1. По каждой выборке вычисляются выборочные средние и выборочные дис-

$$\text{персии: } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \bar{y} = \frac{1}{m} \sum_{i=1}^m y_i, S_y^2 = \frac{1}{m} \sum_{i=1}^m (y_i - \bar{y})^2.$$

$$2. \text{ Вычисляется критериальная статистика } T = \frac{\sqrt{n+m-2}(\bar{x} - \bar{y})}{\sqrt{\frac{n+m}{nm} \sqrt{(n-1)S_x^2 + (m-1)S_y^2}}}.$$

Построение критической области. При условии истинности нулевых гипотез статистика T имеет распределение Стьюдента с $(n + m - 2)$ степенью свободы.

Случай а). Вычисляется квантиль t порядка $1 - \alpha/2$ распределения Стьюдента с $(n + m - 2)$ степенью свободы. Гипотеза H_0 принимается, если выполняется неравенство $|T| \leq t$, иначе гипотеза H_0 отклоняется.

Случай б). Вычисляется квантиль t порядка $1 - \alpha$ распределения Стьюдента с $(n + m - 2)$ степенью свободы. Гипотеза H_0 принимается, если $T \leq t$.

Комментарии

1. Если известна дисперсия совокупностей σ^2 , вместо распределения Стьюдента используют стандартное нормальное распределение, а в формуле вычисления статистики T заменяют значения выборочных дисперсий на σ^2 .
2. Критерий устойчив при умеренных отклонениях распределения выборки от нормального.
3. Критерий также устойчив, если дисперсии генеральных совокупностей незначительно отличаются, а значения n и m приблизительно равны.

Практическая реализация в Excel этого критерия осуществляется с помощью средства Двухвыборочный z-тест с одинаковыми дисперсиями, которое описано в разделе 5.7. Там же приводится пример реализации критерия.

Критерий Беренса–Фишера проверки гипотезы о равенстве математических ожиданий (случай неравных дисперсий)

Этот критерий описан в разделе 2.4.2 и в разделе 5.8, посвященном средству Двухвыборочный t-тест с различными дисперсиями.

Статистическая модель. Выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_m объемом соответственно n и m извлечены из совокупностей, имеющих нормальные распределения с неизвестными дисперсиями σ_1^2 и σ_2^2 и математическими ожиданиями соответственно μ_1 и μ_2 . Равенство дисперсий не предполагается.

Гипотезы

а) Равенство

$$H_0: \mu_1 = \mu_2$$

$$H_1: \mu_1 \neq \mu_2$$

б) Неравенство

$$H_0: \mu_1 \leq \mu_2$$

$$H_1: \mu_1 > \mu_2$$

Задан уровень значимости α .

Вычисления

1. По каждой выборке вычисляются выборочные средние и выборочные дис-

$$\text{персии: } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i, S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \bar{y} = \frac{1}{m} \sum_{i=1}^m y_i, S_y^2 = \frac{1}{m} \sum_{i=1}^m (y_i - \bar{y})^2.$$

2. Вычисляется критериальная статистика $T = \frac{\bar{x} - \bar{y}}{\sqrt{S_x^2/n + S_y^2/m}}$ и дополнитель-

$$\text{но значение } k = \frac{(S_x^2/n + S_y^2/m)^2}{\frac{(S_x^2/n)^2}{n-1} + \frac{(S_y^2/m)^2}{m-1}}.$$

Построение критической области. При условии истинности нулевых гипотез статистика T приближенно имеет распределение Стьюдента с k степенью свободы.

Случай а). Вычисляется квантиль t порядка $1 - \alpha/2$ распределения Стьюдента с k степенью свободы. Гипотеза H_0 принимается, если выполняется неравенство $|T| \leq t$, иначе гипотеза H_0 отклоняется.

Случай б). Вычисляется квантиль t порядка $1 - \alpha$ распределения Стьюдента с k степенью свободы. Гипотеза H_0 принимается, если $T \leq t$.

Комментарии

1. Критерий является приближенным. Если нет оснований предполагать, что дисперсии не равны (критерий проверки равенства дисперсий описан ниже), следует применить точный критерий проверки средних при равных дисперсиях. Если сумма объемов выборок больше 30, вместо распределения Стьюдента можно использовать нормальное распределение.
2. Критерий устойчив при умеренных отклонениях распределения выборки от нормального.
3. Если условия применимости критерия явно не выполняются, следует обратить внимание на непараметрические критерии, описанные в разделе 12.1, которые можно использовать как критерии сравнения математических ожиданий.

Практическая реализация в Excel этого критерия осуществляется с помощью средства Двухвыборочный t-тест с различными дисперсиями, которое описано в разделе 5.8. Там же приводится пример реализации критерия.

Критерий проверки гипотезы о равенстве нескольких математических ожиданий (случай равных дисперсий)

Данный критерий является методом однофакторного дисперсионного анализа (см. раздел 3.5.2).

Статистическая модель. Дано k одномерных независимых выборок объемом соответственно n_1, n_2, \dots, n_k , имеющих нормальное распределение с неизвестны-

ми, но равными дисперсиями. Обозначим $n = \sum_{i=1}^k n_i$.

Гипотезы

$H_0: \mu_1 = \mu_2 = \dots = \mu_k$;

H_1 : математические ожидания $\mu_1, \mu_2, \dots, \mu_k$ различны.

Задан уровень значимости α .

Вычисления

1. По каждой выборке вычисляются выборочные средние и выборочные дисперсии по стандартным формулам $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k, s_1^2, s_2^2, \dots, s_k^2$.

2. Вычисляется общее среднее $\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i \bar{x}_i$ и общая выборочная дисперсия

$$\bar{s}^2 = \frac{1}{n-k} \sum_{i=1}^k (n_i - 1) s_i^2.$$

3. Вычисляется критериальная статистика $T = \frac{1}{\bar{s}^2} \frac{1}{k-1} \sum_{i=1}^k n_i (\bar{x}_i - \bar{x})^2$.

Построение критической области. При условии истинности гипотезы H_0 статистика T имеет F -распределение со степенями свободы $v_1 = k - 1$ и $v_2 = n - k$.

Вычисляется квантиль t порядка $1 - \alpha$ F -распределения со степенями свободы $v_1 = k - 1$ и $v_2 = n - k$. Гипотеза H_0 принимается, если выполняется неравенство $T \leq t$, иначе гипотеза H_0 отклоняется.

Комментарии

1. В случае двух выборок критерий эквивалентен критерию Стьюдента проверки гипотезы о равенстве математических ожиданий (см. выше).
2. Критерий устойчив при умеренных отклонениях распределения выборки от нормального, если выборки достаточного большого объема.
3. Критерий устойчив при умеренных отклонениях от требования равенства дисперсий, если выборки примерно одного объема.
4. Можно проверять другие критерии, например:

$H_0: \mu_1 - \mu_2 = \delta, \mu_2 = \mu_3 = \dots = \mu_k$;

H_1 : нулевая гипотеза неверна.

В этом случае следует вычесть δ из значений первой выборки и затем применить критерий.

5. Если нулевая гипотеза отклоняется, то для определения того, какие средние различаются, необходимо применить критерий множественных сравнений Шеффе, описанный в следующем разделе.
6. Если предположения данного критерия явно не выполняются, следует обратить внимание на непараметрический критерий Краскала-Уоллиса (раздел 12.1.3), который можно использовать как критерий сравнения математических ожиданий.

Практическая реализация

На рис. 12.13 показан рабочий лист Excel, реализующий данный критерий. Все основные формулы, необходимые для вычислений, показаны на этом рисунке. В качестве тестовых выборок выступают три выборки. Первая имеет нормальное распределение с единичным математическим ожиданием и единичной дисперсией, вторая и третья — стандартное нормальное распределение. Диапазоны ячеек, содержащие выборочные значения, названы соответственно Выборка1, Выборка2, Выборка3. Обращаем внимание на применение функции СУММПРОИЗВ (см. раздел 6.1.6), в том числе в формулах массива, что позволяет исключить использование промежуточных вычислений.

E5		A = СУММ(E2:E4)					
	A	B	C	D	E	F	G
1	Выборка1	Выборка2	Выборка3	Объемы: Выборок	Уровень значимости		
2	1,6592368	0,8099772	0,5236163	Выборка1	25	0,05	
3	-0,199411	0,176479	-0,728207	Выборка2	30	Статистика	6,10445886 (=СУММПРОИЗВ(E2:E4;
4	0,6018174	0,2088939	1,2458392	Выборка3	35	Критическое значение	(E7:E9-E10)/2/(2*E15))
5	2,2278614	0,1961983	-0,463486	Всего	90		=ФРАСПОБР(G2,E16:E17)
6	0,643879	0,0374412	0,6502358	Средние		Гипотеза	отклоняется
7	0,6547209	-0,324718	0,3942277	Выборка1	0,92742	=ЕСЛИ(G3<G5,"принимается","отклоняется")	
8	0,5042343	-0,072752	-0,817698	Выборка2	0,17242	=СРЗНАЧ(Выборка2)	
9	-1,18163	1,3691956	1,0260893	Выборка3	-0,0276	=СРЗНАЧ(Выборка3)	
10	0,4390144	-1,330909	-0,691974	Общее	0,30435	=СУММПРОИЗВ(E2:E4;E7:E9)/E5	
11	1,8504157	1,3331994	-0,984059	Дисперсии			
12	-0,265926	-0,807652	0,1204452	Выборка1	0,67103	=ДИСП(Выборка1)	
13	1,035655	-1,090504	-0,052257	Выборка2	1,33749	=ДИСП(Выборка2)	
14	0,6917542	-2,265852	-1,552775	Выборка3	1,33749	=ДИСП(Выборка3)	
15	2,0657815	-0,296167	0,8884746	Общая	1,15364	=СУММПРОИЗВ(E2:E4-1;E12:E14)/(E5-3))	
16	0,3022682	-1,487918	-0,574985	v1	2		
17	0,9923062	0,9672316	-1,158917	v2	87	=E5-3	
18	1,1692431	1,3523859	0,2537612				
19	1,4007521	0,5085407	-0,30082				

Рис. 12.13. Критерий проверки гипотезы о равенстве нескольких математических ожиданий

Этот критерий реализует также средство Однофакторный дисперсионный анализ, описанное в разделе 5.11. Там же приводится пример использования этого средства.

Критерий множественных сравнений Шеффе

Если предыдущий критерий сравнения математических ожиданий отвергает нулевую гипотезу о равенстве всех математических ожиданий, то критерий множественных сравнений Шеффе позволяет определить, математические ожидания каких выборок выделяются из общего ряда. С помощью этого метода можно провести несколько парных сравнений выборок, не увеличивая при этом вероятность ошибки первого рода.

Статистическая модель. Дано k одномерных независимых выборок объемом соответственно n_1, n_2, \dots, n_k , имеющих нормальное распределение с неизвестными, но равными дисперсиями. Обозначим $n = \sum_{i=1}^k n_i$.

Гипотезы

H_0 : $c_1\mu_1 + c_2\mu_2 + \dots + c_k\mu_k = 0$, где c_1, c_2, \dots, c_k — заданные постоянные, сумма которых равна нулю;

H_1 : нулевая гипотеза неверна.

Задан уровень значимости α .

Вычисления

1. По каждой выборке вычисляются выборочные средние и выборочные дисперсии по стандартным формулам $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k, s_1^2, s_2^2, \dots, s_k^2$.

2. Вычисляется общее среднее $\bar{x} = \frac{1}{n} \sum_{i=1}^k n_i \bar{x}_i$ и общая выборочная дисперсия

$$\bar{s}^2 = \frac{1}{n-k} \sum_{i=1}^k (n_i - 1) s_i^2.$$

3. Вычисляется критериальная статистика $T = \frac{\left(\sum_{i=1}^k c_i \bar{x}_i \right)^2}{\bar{s}^2 (k-1) \sum_{i=1}^k c_i^2 / n_i}.$

Построение критической области. При условии истинности гипотезы H_0 статистика T имеет F -распределение со степенями свободы $v_1 = k - 1$ и $v_2 = n - k$.

Вычисляется квантиль t порядка $1 - \alpha$ F -распределения со степенями свободы $v_1 = k - 1$ и $v_2 = n - k$. Гипотеза H_0 принимается, если выполняется неравенство $T \leq t$, иначе гипотеза H_0 отклоняется.

Комментарии

1. Критерий часто применяют для серии сравнений типа

$$H_0: \mu_1 - \mu_2 = 0;$$

$$H_1: \mu_1 - \mu_2 \neq 0.$$

При попарных сравнениях обычно сначала сравнивают выборку с наименьшим выборочным средним с каждой последующей, отмечая те выборки, для которых критерий Шеффе не отвергает нулевую гипотезу. Затем повторяют сравнения с выборкой, имеющей второе по величине значение выборочного среднего, и т.д.

2. Критерий устойчив при умеренных отклонениях от требования равенства дисперсий, если выборки примерно одного объема.

Практическая реализация

Реализация этого критерия в Excel незначительно отличается от реализации предыдущего критерия. На рис. 12.14 показан рабочий лист, реализующий критерий Шеффе. Кроме формулы вычисления критериальной статистики в ячейке G7 (формула приведена на рисунке), все остальные формулы совпадают с аналогичными формулами, показанными на рис. 12.13. В качестве тестовых данных также используются выборки из предыдущего раздела. На рис. 12.14 показано сравнение второй и третьей выборок — различие между ними незначимо. На рис. 12.15 сравниваются первая и третья выборки, здесь различие значимо — нулевая гипотеза отклоняется.

12.3.2. Проверка гипотез о дисперсиях нормальных распределений

В этом разделе приведены критерии проверки равенства дисперсий нормальных совокупностей.

	A	B	C	D	E	F	G	H	I
1	Выборка1	Выборка2	Выборка3	Объемы	выборки	Уровень	значимости		
2	1,6592368	0,8098772	0,5236163	Выборка1	25		0,05		
3	-0,198411	0,176478	-0,728207	Выборка2	30	Коэффициенты ci			
4	0,6018174	0,2068939	1,2459392	Выборка3	35	c1	0		
5	2,2278614	0,1961983	-0,463486	Всего	90	c2	1		
6	0,643879	0,0374412	0,6502358	Средние		c3	-1		
7	0,6547209	-0,324718	0,3942277	Выборка1	0,92742	Статистика	0,280158		
8	0,5042343	-0,072752	-0,817698	Выборка2	0,17242	Критическое значение	3,101292		
9	-1,18163	1,3891956	1,0260893	Выборка3	-0,0276				
10	0,4390144	-1,330909	-0,681974	Общее	0,30435	Гипотеза принимается			
11	1,8504157	1,3331994	-0,984059	Дисперсии					
12	-0,265926	-0,807652	0,1204452	Выборка1	0,67103	{=(СУММПРОИЗВ(E7:E9,G4:G6)^2)/(E16^E15*СУММ(((G4:G6)^2)/E2:E4))}			
13	1,035655	-1,090504	-0,052257	Выборка2	1,33749				
14	0,6917542	-2,265952	-1,552775	Выборка3	1,33749				
15	2,0657815	-0,296167	0,8984746	Общая	1,15364				
16	0,3022682	-1,487918	-0,574985	v1	2				
17	0,9923062	0,9672316	-1,158917	v2	87				
18	1,1692431	1,3523659	0,2537812						

Рис. 12.14. Критерий множественных сравнений Шеффе: сравнение второй и третьей выборок

	A	B	C	D	E	F	G	H
1	Выборка1	Выборка2	Выборка3	Объемы	выборки	Уровень	значимости	
2	1,6592368	0,8098772	0,5236163	Выборка1	25		0,05	
3	-0,198411	0,176478	-0,728207	Выборка2	30	Коэффициенты ci		
4	0,6018174	0,2068939	1,2459392	Выборка3	35	c1	1	
5	2,2278614	0,1961983	-0,463486	Всего	90	c2	0	
6	0,643879	0,0374412	0,6502358	Средние		c3	-1	
7	0,6547209	-0,324718	0,3942277	Выборка1	0,92742	Статистика	5,784993	
8	0,5042343	-0,072752	-0,817698	Выборка2	0,17242	Критическое значение		
9	-1,18163	1,3891956	1,0260893	Выборка3	-0,0276		3,101292	
10	0,4390144	-1,330909	-0,681974	Общее	0,30435	Гипотеза отклоняется		
11	1,8504157	1,3331994	-0,984059	Дисперсии				
12	-0,265926	-0,807652	0,1204452	Выборка1	0,67103			
13	1,035655	-1,090504	-0,052257	Выборка2	1,33749			
14	0,6917542	-2,265952	-1,552775	Выборка3	1,33749			
15	2,0657815	-0,296167	0,8984746	Общая	1,15364			
16	0,3022682	-1,487918	-0,574985	v1	2			
17	0,9923062	0,9672316	-1,158917	v2	87			
18	1,1692431	1,3523659	0,2537812					

Рис. 12.15. Сравнение первой и третьей выборок

Критерий Фишера проверки равенства дисперсий

Этот критерий описан в разделе 2.4.2 и в разделе 5.10, посвященном средству Двухвыборочный F-тест для дисперсий из пакета анализа.

Статистическая модель. Выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_m объемом соответственно n и m извлечены из совокупностей, имеющих нормальные распределения с неизвестными дисперсиями σ_1^2 и σ_2^2 и математическими ожиданиями μ_1 и μ_2 соответственно.

Гипотезы

а) Равенство

$$H_0: \sigma_1^2 = \sigma_2^2$$

$$H_1: \sigma_1^2 \neq \sigma_2^2$$

б) Неравенство

$$H_0: \sigma_1^2 \leq \sigma_2^2$$

$$H_1: \sigma_1^2 > \sigma_2^2$$

Задан уровень значимости α .

Вычисления

1. Для каждой выборки вычисляются выборочные дисперсии

$$S_x^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2, \quad S_y^2 = \frac{1}{m} \sum_{i=1}^m (y_i - \bar{y})^2.$$

2. В качестве критериальной статистики вычисляется дисперсионное отношение Фишера $F = \frac{S_x^2}{S_y^2}$.

Построение критической области. В случае истинности нулевой гипотезы статистика F имеет F -распределение со степенями свободы k_1 и k_2 , где $k_1 = n - 1$, $k_2 = m - 1$, если $F \geq 1$, и $k_1 = m - 1$, $k_2 = n - 1$, если $F < 1$.

Случай а). Вычисляется критическое значение $t_{кр}$ как квантиль порядка $1 - \alpha/2$ F -распределения со степенями свободы k_1 и k_2 . Гипотеза H_0 принимается, если выполняется неравенство $F \leq t_{кр}$, иначе гипотеза H_0 отклоняется.

Случай б). Вычисляется критическое значение $t_{кр}$ как квантиль порядка $1 - \alpha$ F -распределения со степенями свободы k_1 и k_2 . Гипотеза H_0 принимается, если $F \leq t_{кр}$.

Комментарии

1. Если известны математические ожидания выборок, то в формулах вычисления выборочных дисперсий средние выборок заменяются известными значениями математических ожиданий. В этом случае критические значения вычисляются как квантили F -распределения со степенями свободы k_1 и k_2 , где $k_1 = n$, $k_2 = m$, если $F \geq 1$, и $k_1 = m$, $k_2 = n$, если $F < 1$.
2. Критерий неустойчив при отклонении от нормальности.

Практическая реализация в Excel этого критерия осуществляется с помощью средства Двухвыборочный F-тест для дисперсий, которое описано в разделе 5.10. Там же приводится пример реализации критерия.

Критерий Бартлета проверки равенства нескольких дисперсий

Статистическая модель. Дано k одномерных независимых выборок объемом соответственно n_1, n_2, \dots, n_k , имеющих нормальное распределение с дисперсиями

$$\sigma_1^2, \sigma_2^2, \dots, \sigma_k^2. \text{ Обозначим } n = \sum_{i=1}^k n_i.$$

Гипотезы

$$H_0: \sigma_1^2 = \sigma_2^2 = \dots = \sigma_k^2;$$

H_1 : дисперсии различны.

Задан уровень значимости α .

Вычисления

1. По каждой выборке вычисляются выборочные средние и выборочные дисперсии по стандартным формулам $\bar{x}_1, \bar{x}_2, \dots, \bar{x}_k, s_1^2, s_2^2, \dots, s_k^2$.
2. Вычисляются величины $C = 1 + \frac{1}{3(k-1)} \left(\sum_{i=1}^k \frac{1}{n_i - 1} - \frac{1}{n - k} \right)$ и $S^2 = \frac{1}{n - k} \sum_{i=1}^k (n_i - 1) s_i^2$.
3. Вычисляется критерияльная статистика $T = \frac{1}{C} \sum_{i=1}^k (n_i - 1) \ln \left(\frac{S^2}{s_i^2} \right)$.

Построение критической области. При условии истинности гипотезы H_0 статистика T приближенно имеет распределение χ^2 с $(k - 1)$ степенью свободы.

Вычисляется квантиль t порядка $1 - \alpha$ распределения χ^2 с $(k - 1)$ степенью свободы. Гипотеза H_0 принимается, если выполняется неравенство $T \leq t$, иначе гипотеза H_0 отклоняется.

Комментарии

1. Критерий является приближенным. В случае двух выборок следует применять точный критерий Фишера.
2. Критерий очень чувствителен к отклонениям распределения выборок от нормального.

Практическая реализация

На рис. 12.16 показан рабочий лист Excel, реализующий данный критерий. Все основные формулы, необходимые для вычислений, также показаны на этом рисунке. В качестве тестовых выборок выступают три выборки, имеющие стандартное нормальное распределение. Диапазоны ячеек, содержащие выборочные значения, названы соответственно Выборка1, Выборка2, Выборка3. Обращаем внимание на применение формул массивов, что дает возможность исключить промежуточные вычисления.

E5		F5		G5		H5		I5	
A		B		C		D		E	
1	Выборка1	Выборка2	Выборка3	Объемы	выборки	Уровень	значимости		
2	-0,697373	0,2498028	-0,386624	Выборка1	20		0,05		
3	0,8938723	-0,062458	-1,003505	Выборка2	30	Статистика	2,471478	=СУММПРОИЗВ((E2:E4-1);	
4	-3,56E-05	-1,651959	1,0244357	Выборка3	35	Критическое значение		LN(E15/E12:E14)/E16}	
5	0,1811017	-1,244861	-0,208838	Всего	85		5,93147636	=ХИ2ОБР(G2;2)	
6	-0,227035	-1,679814	-0,280816	Средние		Гипотеза	принимается		
7	0,7301227	1,0627774	0,7174612	Выборка1	0,01432	=ЕСЛИ(G3<G5,"принимается","отклоняется")			
8	-0,577312	0,1562593	-1,249709	Выборка2	0,09968	=СРЗНАЧ(Выборка2)			
9	-0,248306	-0,113105	-0,465794	Выборка3	0,05168	=СРЗНАЧ(Выборка3)			
10	0,6751592	0,0176725	0,4148023	Общее	0,05983	=СУММПРОИЗВ(E2:E4;E7:E9)/E5			
11	-1,636537	1,635215	-0,039106	Дисперсии					
12	0,6516314	-0,465149	0,0630081	Выборка1	0,68717	=ДИСП(Выборка1)			
13	-1,423338	1,245734	-0,213721	Выборка2	1,26077	=ДИСП(Выборка2)			
14	0,1591798	1,6592997	0,4679564	Выборка3	1,29077	=ДИСП(Выборка3)			
15	0,6916627	0,3077727	-0,263893	Общая	1,14323	=СУММПРОИЗВ(E2:E4-1;E12:E14)/(E5-3)			
16	-0,890302	0,834256	0,3868295	С	1,01738	=(1+(СУММ(1/(E2:E4-1))-1/(E5-3))/(3*2))			
17	-0,311894	0,5197851	-0,389125						
18	-0,72354	-1,115367	0,6231784						

Рис. 12.16. Критерий проверки гипотезы о равенстве нескольких дисперсий

12.3.3. Непараметрический критерий Ансари–Бредли проверки гипотезы о равенстве дисперсий

Этот критерий используется тогда, когда не выполняется предположение о нормальности распределений выборок.

Статистическая модель. Выборки x_1, x_2, \dots, x_n и y_1, y_2, \dots, y_m объемом соответственно n и m извлечены из совокупностей, имеющих непрерывные распределения с неизвестными дисперсиями σ_1^2 и σ_2^2 . Предполагается, что распределения имеют одинаковые медианы. Также предполагается, что $N = n + m \geq 20$.

Гипотезы

$$H_0: \sigma_1^2 = \sigma_2^2;$$

$$H_1: \sigma_1^2 \neq \sigma_2^2.$$

Задан уровень значимости α .

Вычисления

1. Обе выборки объединяются в единую выборку, и по объединенной выборке строится вариационный ряд $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(N)}$.
2. Вычисляются ранги выборочных значений следующим образом. Наибольшему и наименьшему значениям в объединенной выборке присваивается ранг 1. Следующим по величине наименьшим и наибольшим значениям присваивается ранг 2 и т.д. Если встречаются одинаковые значения, то им приписываются равные средние ранги.
3. Для одной из выборок подсчитывается сумма рангов R , которые получили ее выборочные значения в объединенной выборке. Пусть для определенности подсчитывается сумма рангов первой выборки.
4. Вычисляется критериальная статистика

$$T = \frac{R - \frac{1}{4}n(N+2)}{\sqrt{\frac{mn(N+2)(N-2)}{48(N-1)}}}, \text{ если } N - \text{четное число, и}$$

$$T = \frac{R - \frac{n(N+1)^2}{4N}}{\sqrt{\frac{mn(N+1)(3+N^2)}{48N^2}}}, \text{ если } N - \text{нечетное число.}$$

(Если вычисляется сумма рангов второй выборки, то в этих формулах в числителе n заменяется на m .)

Построение критической области. При условии истинности гипотезы H_0 статистика T приближенно имеет стандартное нормальное распределение.

Вычисляется квантиль t порядка $1 - \alpha/2$ стандартного нормального распределения. Гипотеза H_0 принимается, если выполняется неравенство $|T| \leq t$, иначе гипотеза H_0 отклоняется.

Комментарии

1. Критерий является асимптотическим (отсюда требование, чтобы $N = n + m \geq 20$). При малых N в качестве критериальной статистики ис-

пользуется величина R , имеющая специальное распределение, квантили которого используются для построения критической области [22].

2. В случае, если есть совпадающие значения, статистику T можно вычислить по более сложным формулам, учитывающим эти совпадения [13].
3. Если распределения имеют различные медианы и значения этих медиан известны (или хотя бы известны их оценки), этот метод также можно применять, если вместо исходных выборок использовать выборки, значения которых уменьшены на величины медиан (для каждой выборки — своя медиана).

Практическая реализация

На рис. 12.17 показан рабочий лист Excel, реализующий данный критерий. Все основные формулы, необходимые для вычислений, также показаны на этом рисунке. В качестве тестовых выборок взяты две выборки, имеющие равномерное распределение: первая на интервале $[-1, 1]$, вторая — на интервале $[-2, 2]$. Диапазоны ячеек, содержащие выборочные значения, названы соответственно **Выборка1** и **Выборка2**. В столбцах C и D с помощью формулы массивов

{=ЕСЛИ(РАНГ(A2:B26;A2:B26;1)>ЦЕЛОЕ(\$F\$4/2);\$F\$4-
РАНГ(A2:B26;A2:B26;1)+1;РАНГ(A2:B26;A2:B26;1))}

вычисляются ранги выборочных значений, как описано выше, в п. 2. Здесь функция $\text{РАНГ}(A2:B26;A2:B26;1)$ строит виртуальный массив “стандартных” рангов объединенной выборки. Если значения этих рангов R_i превышают половину числа N общего объема выборок, то они замещаются значением $N - R_i + 1$. В противном случае значения рангов остаются без изменения. Обращаем внимание, что значение статистики T вычисляется по разным формулам в ячейках H4 и H5 для четного и нечетного значений N . Конечно, эти формулы можно объединить в одну с помощью функции ЕСЛИ, однако такая объединенная формула будет весьма сложной и нечитаемой. Выбор значения T , в зависимости от четности или нечетности N , осуществляется в формуле ячейки G8 при определении, отвергается нулевая гипотеза или принимается. В данном случае гипотеза о равенстве дисперсий отвергается.

F2		=СЧЕТ(Выборка1)					
	A	B	C	D	E	F	G
1	Выборка1	Выборка2	Ранг1	Ранг2	Объемы выборок	Уровень значимости	
2	-0,818437	-1,234471	7	6	Выборка1	20	0,05
3	0,551584	1,1024277	16	5	Выборка2	25	
4	-0,475429	-0,21194	14	19	Всего	45	Статистика
5	-0,765888	-0,616445	9	12	R	296	N нечетно
6	0,633356	-1,263819	12	5	=СУММ(C2:C21)	Критическое значение	=(F5-(F2*(F4+2)/4))КОРЕНЬ(F2*
7	0,610766	0,8795388	14	8			F3*(F4+2)*(F4-2)/(48*(F4-1)))
8	0,75859	0,9752064	10	7	Гипотеза отвергается		=(F5-(F2*((F4+1)*2)/(4*F4)))/
9	0,939609	-1,652792	17	3			КОРЕНЬ(F2*F3*(F4+1)*
10	-0,112189	0,7730564	21	9	=ЕСЛИ(ЕСЛИ(ЧЕТН(F4),H4,H5)<H7,"принимается","отклоняется")		(F4*F4+3)/(48*F4*F4))
11	0,611155	1,0323468	13	6			=НОРМСТОБР(1-H2/2)
12	0,574476	-0,476639	15	13	В диапазоне C2:D26 формула массива:		
13	0,353518	-0,22228	18	10	{=ЕСЛИ(РАНГ(A2:B26;A2:B26;1)>ЦЕЛОЕ(\$F\$4/2);\$F\$4-		
14	0,323263	1,8310513	20	1	РАНГ(A2:B26;A2:B26;1)+1;РАНГ(A2:B26;A2:B26;1))}		
15	-0,815433	-0,399453	8	15			
16	0,050796	1,8156326	23	2			
17	-0,707337	0,1733967	10	21			
18	-0,338365	-0,225504	16	17			

Рис. 12.17. Критерий Ансари-Бредли проверки гипотезы о равенстве дисперсий

12.3.4. Проверка гипотез о равенстве биномиальных вероятностей

Статистическая модель. Имеются две серии наблюдений за экспериментом. В первой серии в каждом эксперименте с вероятностью p_1 происходит событие "1" ("успех"), во второй серии это событие происходит с вероятностью p_2 . Пусть в первой серии зафиксировано n экспериментов, из них в r_1 случаях наблюдалось событие "1". Во второй серии зафиксировано m экспериментов, из них в r_2 случаях наблюдалось событие "1". Размеры серий больше 20.

Гипотезы

а) Равенство

б) Неравенство

$$H_0: p_1 - p_2 = \delta$$

$$H_0: p_1 - p_2 \leq \delta$$

$$H_1: p_1 - p_2 \neq \delta$$

$$H_1: p_1 - p_2 > \delta$$

Задан уровень значимости α .

Вычисления

1. Вычисляются точечные оценки вероятностей p_1 и p_2 : $\hat{p}_1 = r_1/n$, $\hat{p}_2 = r_2/m$

$$\text{и дополнительно } \hat{\delta} = \hat{p}_1 - \hat{p}_2, A_{n,m} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n} + \frac{\hat{p}_2(1-\hat{p}_2)}{m}}.$$

2. Вычисляется критериальная статистика $T = \frac{\hat{\delta} - \delta}{A_{n,m}}$.

Построение критической области. При условии истинности нулевых гипотез статистика T приближенно имеет стандартное нормальное распределение.

Случай а). Вычисляется квантиль t порядка $1 - \alpha/2$ стандартного нормального распределения. Гипотеза H_0 принимается, если выполняется неравенство $|T| \leq t$, иначе гипотеза H_0 отклоняется.

Случай б). Вычисляется квантиль t порядка $1 - \alpha$ стандартного нормального распределения. Гипотеза H_0 принимается, если выполняется неравенство $T \leq t$, иначе гипотеза H_0 отклоняется.

Комментарии

1. Критерий является приближенным и основывается на аппроксимации биномиального распределения нормальным. Отсюда требование, чтобы объемы выборок были не меньше 20.
2. Критерий построен на основе доверительного интервала для разностей биномиальных вероятностей (см. раздел 12.2.4).

Практическая реализация в Excel данного метода не вызывает затруднений.

Статистический анализ зависимостей

В этой части...

Глава 13. Корреляционный анализ

Глава 14. Сравнение зависимых выборок

Глава 15. Регрессионный анализ

В этой части описаны методы анализа статистических зависимостей, включающие в себя широкий спектр статистических алгоритмов. В главе 13 рассмотрены методы корреляционного анализа, которые устанавливают сам факт статистической зависимости между данными, а также способы построения доверительных интервалов и критерии проверки гипотез о значениях коэффициента корреляции. В главе 14 показаны методы сравнения параметров распределений зависимых компонентов многомерных выборок. В главе 15 описан круг задач, связанных с построением регрессий, начиная с общей вычислительной схемы определения коэффициентов уравнений регрессии и заканчивая критериями проверки адекватности построенного уравнения регрессии.

Корреляционный анализ

Настоящая глава посвящена задаче установления самого факта наличия статистически значимой связи между переменными. В общем виде эта задача описана в главе 3. Напомним, что методы, применяемые для ее решения, зависят от природы исследуемых случайных переменных (количественные, порядковые или классификационные), от выбранного показателя статистической зависимости (индекс или коэффициент корреляции, ранговый коэффициент корреляции и т.п.) и от конкретной решаемой задачи (точечное или интервальное оценивание показателя статистической зависимости, проверка гипотезы о значении показателя статистической зависимости).

13.1. Критерии независимости

В этом разделе описаны критерии проверки гипотез о независимости многомерных случайных величин. Для количественных случайных величин это критерии проверки гипотез о нулевом значении коэффициента корреляции, для порядковых случайных величин аналогичные критерии строятся на основе ранговых коэффициентов корреляции, для классификационных величин применяется анализ таблиц сопряженности (см. главу 3). Для всех этих методов справедливо "правило вложенности" — методы, применимые для классификационных случайных величин, также применимы для порядковых и количественных случайных величин; методы, применимые для порядковых случайных величин, также применимы для количественных величин. Однако по возможности следует использовать критерии, предназначенные для конкретного типа случайных переменных. Исключение составляют критерии, построенные на основе ранговых коэффициентов корреляции, поскольку критерий независимости для количественных случайных величин (см. следующий раздел) является приближенным и будет точным только для нормально распределенных величин. Кроме того, непараметрический критерий на основе коэффициента конкордации позволяет оценить взаимозависимость нескольких (больше двух) случайных величин.

Также необходимо помнить, что критерии, построенные на основе коэффициента корреляции, *не доказывают независимость* случайных величин: незначимое отличие коэффициента корреляции от нуля говорит только о том, что *отсутствует линейная зависимость* между случайными величинами. Возможны виды нелинейной зависимости между случайными величинами, когда коэффициент корреляции между ними равен нулю. Но на практике, как правило, некоррелируемость отождествляют с независимостью.

13.1.1. Критерий независимости на основе преобразования Фишера

Нормализующее z -преобразование Фишера и его свойства описаны в разделе 3.3.1.

Статистическая модель. Выборочные значения $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ являются реализацией случайной величины $Z = (X, Y)$, имеющей произвольное двумерное распределение с конечными моментами второго порядка и с коэффициентом корреляции ρ . Объем выборки — не менее 20.

Гипотезы

H_0 : коэффициент корреляции $\rho = 0$;

H_1 : коэффициент корреляции $\rho \neq 0$.

Задан уровень значимости α .

Вычисления

1. Вычисляются выборочные средние $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ и $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

2. Вычисляется точечная оценка коэффициента корреляции

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

3. Вычисляется критериальная статистика $T = \frac{\sqrt{n-3}}{2} \ln \frac{1+r}{1-r}$.

Построение критической области. При условии истинности нулевой гипотезы статистика T асимптотически имеет стандартное нормальное распределение.

Вычисляется критическое значение t как квантиль порядка $1 - \alpha/2$ стандартного нормального распределения. Гипотеза H_0 принимается, если выполняется неравенство $|T| \leq t$, иначе гипотеза H_0 отклоняется.

Комментарии

1. Этот критерий приближенный. Точное распределение статистики T зависит от распределения случайной величины Z ; оно весьма сложное.

2. Для нормально распределенной случайной величины существует точный критерий (см. следующий раздел).

3. Преобразование Фишера $z = \frac{1}{2} \ln \frac{1+r}{1-r}$, являющееся основой данного критерия,

только асимптотически имеет нормальное распределение. Отсюда требование, чтобы объем выборки был не менее 20. Точность аппроксимации распределения величины z нормальным распределением зависит от распределения случайной величины Z . Поэтому точность может быть различной для разных выборочных распределений.

4. Во многих случаях, когда распределение случайной величины Z очень далеко от нормального или дискретно, следует применять непараметрические критерии независимости, описанные ниже.

Практическая реализация в Excel данного критерия не вызывает затруднений. Отметим, что для вычисления выборочного коэффициента корреляции в Excel предусмотрена функция КОРРЕЛ, а преобразование Фишера вычисляет функция ФИШЕР.

13.1.2. Критерий независимости для двумерных нормальных совокупностей

Статистическая модель. Выборочные значения $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ являются реализацией случайной величины $Z = (X, Y)$, имеющей двумерное нормальное распределение с коэффициентом корреляции ρ .

Гипотезы

H_0 : коэффициент корреляции $\rho = 0$;

H_1 : коэффициент корреляции $\rho \neq 0$.

Задан уровень значимости α .

Вычисления

1. Вычисляются выборочные средние $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ и $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

2. Вычисляется точечная оценка коэффициента корреляции

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

3. Вычисляется критерияльная статистика $T = \frac{r\sqrt{n-2}}{\sqrt{1-r^2}}$.

Построение критической области. При условии истинности нулевой гипотезы статистика T имеет распределение Стьюдента со степенью свободы $(n - 2)$.

вычисляется критическое значение t как квантиль порядка $1 - \alpha/2$ распределения Стьюдента со степенью свободы $(n - 2)$. Гипотеза H_0 принимается, если выполняется неравенство $|T| \leq t$; иначе гипотеза H_0 отклоняется.

Комментарии

1. Этот критерий можно применять при умеренных отклонениях выборочного распределения от нормального.
2. В разделе 13.3.1 приведен критерий проверки значения коэффициента корреляции.

Практическая реализация в Excel данного критерия не вызывает затруднений. Для вычисления выборочного коэффициента корреляции в Excel предусмотрена функция КОРРЕЛ.

13.1.3. Критерий независимости на основе рангового коэффициента корреляции Спирмена

Ранговый коэффициент корреляции Спирмена и его свойства описаны в разделе 3.3.2.

Статистическая модель. Выборочные значения $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ являются реализацией случайной величины $Z = (X, Y)$, имеющей произвольное двумерное распределение с конечными моментами второго порядка. Объем выборки — более 10.

Гипотезы

H_0 : ранговый коэффициент корреляции Спирмена $r_s = 0$;

H_1 : ранговый коэффициент корреляции $r_s \neq 0$.

Задан уровень значимости α .

Вычисления

1. Каждому выборочному значению (x_i, y_i) присваиваются ранги (r_i, q_i) путем построения отдельных вариационных рядов $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ и $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$. Если есть совпадающие выборочные значения, то этим значениям присваиваются одинаковые ранги, равные среднему рангов, которые были бы им присвоены при отсутствии равенства значений.

2. Вычисляется ранговый коэффициент корреляции Спирмена

$$r_s = 1 - \frac{6}{n^3 - n} \sum_{i=1}^n (r_i - q_i)^2.$$

3. Вычисляется критериальная статистика $T = \frac{r_s \sqrt{n-2}}{\sqrt{1-r_s^2}}$.

Построение критической области. При условии истинности нулевой гипотезы коэффициент r_s имеет специальное распределение; статистика T асимптотически имеет распределение Стьюдента со степенью свободы $(n - 2)$.

Вычисляется критическое значение t как квантиль порядка $1 - \alpha/2$ распределения Стьюдента со степенью свободы $(n - 2)$. Гипотеза H_0 принимается, если выполняется неравенство $|T| \leq t$, иначе гипотеза H_0 отклоняется.

Комментарии

1. Этот критерий приближенный и применяется для выборок объемом не менее 10. Для малых выборок в качестве критериальной статистики берется коэффициент r_s , а критическое значение определяется из таблицы специального распределения Спирмена.

2. Если нет достаточных оснований отвергнуть гипотезу о нормальном распределении генеральной совокупности, из которой извлекается выборка, то целесообразно применять более мощный метод, основанный на коэффициенте корреляции (см. предыдущий раздел).

Практическая реализация

На рис. 13.1 показан рабочий лист Excel, реализующий данный критерий. В столбцах A и B записаны выборочные значения (в данном случае значения независимы и имеют равномерные распределения на интервале $[0, 1]$). Диапазон

$$\{=СУММКВРАЗН(РАНГ(X;X;1);РАНГ(Y;Y;1))\}$$

	D1	B	=СЧЕТ(X)					
	A	B	C	D	E	F	G	H
1	X	Y	Объем	20	Уровень значимости	0,05		
2	0,187665	0,019115	Сумма кв. разностей рангов					
3	0,270794	0,492598		1566	{=СУММКВРАЗН(РАНГ(X;X;1),РАНГ(Y;Y;1))}			
4	0,678216	0,451612	Коэффициент Спирмена					
5	0,973084	0,289608		-0,17744	=1-6*D3/(D1^3-D1)			
6	0,874119	0,694154	Критериальная статистика					
7	0,938301	0,355918		-0,76497	=D5*КОРЕНЬ(D1-2)/КОРЕНЬ(1-D5^D5)			
8	0,260673	0,804617	Критическое значение					
9	0,2856	0,336788		2,445004	=СТЮДРАСПОБР(G1/2,D1-2)			
10	0,086681	0,828136	Гипотеза					
11	0,506678	0,877876			=ЕСЛИ(ABS(D7)<D9,"принимается","отклоняется")			
12	0,29262	0,612735						
13	0,03465	0,541228						
14	0,512648	0,166106						
15	0,923784	0,497093						
16	0,690115	0,970908						

13.1.4. Критерий независимости на основе рангового коэффициента корреляции Кендалла

Статистическая модель. Выборочные значения $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ являются реализацией случайной величины $Z = (X, Y)$, имеющей произвольное двумерное распределение с конечными моментами второго порядка. Объем выборки — более 10.

H_0 : ранговый коэффициент корреляции Кендалла $r_K = 0$;

$$H_1: \text{ранговый коэффициент корреляции } r_g \neq 0.$$

386 Часть IV. Статистический анализ зависимостей

1. Каждому выборочному значению (x_i, y_i) присваиваются ранги (r_i, q_i) путем построения отдельных вариационных рядов $x_{(1)} \leq x_{(2)} \leq \dots \leq x_{(n)}$ и $y_{(1)} \leq y_{(2)} \leq \dots \leq y_{(n)}$. Если есть совпадающие выборочные значения, то этим значениям присваиваются одинаковые ранги, равные среднему рангов, которые были бы им присвоены при отсутствии равенства значений.

2. Полученная последовательность рангов $(r_1, q_1), (r_2, q_2), \dots, (r_n, q_n)$ упорядочивается по возрастанию рангов r_i — получается последовательность $(1, q_{(1)}), (2, q_{(2)}), \dots, (n, q_{(n)})$.

3. Вычисляется ранговый коэффициент корреляции Кендалла

$$r_K = \frac{2}{n(n-1)} \sum_{i=1}^n \sum_{j=i+1}^n \text{sign}(q_{(j)} - q_{(i)}), \text{ где функция } \text{sign}(x) \text{ принимает значение } +1, \text{ если } x > 0, \text{ и значение } -1, \text{ если } x < 0.$$

4. Вычисляется критериальная статистика $T = r_K \sqrt{\frac{9n(n-1)}{2(2n+5)}}$.

Построение критической области. При условии истинности нулевой гипотезы коэффициент r_K имеет специальное распределение Кендалла, статистика T асимптотически имеет стандартное нормальное распределение.

Определяется критическое значение t как квантиль порядка $1 - \alpha/2$ стандартного нормального распределения. Гипотеза H_0 принимается, если выполняется неравенство $|T| \leq t$, иначе гипотеза H_0 отклоняется.

Комментарии

1. Этот критерий приближенный и применяется для выборок объемом не менее 10. Для малых выборок в качестве критериальной статистики берется коэффициент r_K , а критическое значение определяется из таблицы специального распределения Кендалла.
2. Если нет достаточных оснований отвергать гипотезу о нормальном распределении генеральной совокупности, из которой извлекается выборка, то целесообразно применять более мощный метод из раздела 13.1.2.

Практическая реализация

На рис. 13.2 показан рабочий лист Excel, реализующий данный критерий. В столбцах A и B записаны выборочные значения (как и в предыдущем примере, выборочные значения независимы и имеют равномерные распределения на интервале $[0, 1]$). Диапазон ячеек в столбце A, содержащий выборочные значения, назван X, а соответствующий диапазон в столбце B назван Y. Все формулы, необходимые для вычисления критерия, показаны на рис. 13.2.

К сожалению, для реализации критерия некоторые действия необходимо выполнить вручную и невозможно обойтись без некоторых промежуточных вычислений. В столбцах A и B записаны выборочные значения, а в столбцах C и D подсчитаны ранги этих значений с помощью формул массивов $\{=\text{РАНГ}(X;X;1)\}$ и $\{=\text{РАНГ}(Y;Y;1)\}$. (Диапазон ячеек в столбце A, содержащий выборочные значения, назван X, а соответствующий диапазон в столбце B назван Y.) Далее необходимо сортировать двухстолбцовый диапазон, содержащий ранги, по значениям рангов

столбца С. Это можно сделать в тех же столбцах С и D, предварительно преобразовав формулы, по которым вычислялись ранги, в значения. Для этого надо выделить диапазон C2:D21, содержащий ранги, скопировать его и, не отменяя его выделение, выполнить команду Правка⇒Специальная вставка. В открывшемся одноименном диалоговом окне следует установить переключатель Значения и щелкнуть на кнопке ОК. На рис. 13.2 для наглядности ранги сортируются в соседних столбцах Е и F, в которые они скопированы из диапазона C2:D21 как значения (опять с помощью диалогового окна Специальная вставка). Сортировка осуществляется с помощью команды Данные⇒Сортировка. В столбце G реализуется часть формулы $\sum_{j=i+1}^n \text{sign}(q_{(j)} - q_{(i)})$ вычисления коэффициента Кендалла. Для

этого в ячейку G2 введена формула массива {=СУММ(ЗНАК(F3:\$F\$21-F2))}, которая сначала создает виртуальный массив значений $\text{sign}(q_{(j)} - q_{(i)})$ ($j \geq 2$), а затем суммирует эти значения. Функция ЗНАК — это эквивалент функции sign. Обращаем внимание, что в этой формуле используются относительные ссылки на ячейку F2 и на начало диапазона F3:\$F\$21. Данная формула копируется вниз до ячейки F20, при этом адреса ячеек F2 и F3 соответствующим образом модифицируются, а конечная ячейка \$F\$21 диапазона суммирования остается неизменной. Диапазон ячеек в столбце G назван Знаки, это имя используется в формуле ячейки I3. Остальные формулы данного рабочего листа очевидны.

G2 {=СУММ(ЗНАК(F3:\$F\$21-F2))}												
	A	B	C	D	E	F	G	H	I	J	K	L
1	X	Y	Ранги	Сортировка	Знаки	Объем	20	Уровень значимости				
2	0,147	0,555	4	10	1	17	-13	Сумма знаков			0,05	
3	0,092	0,272	2	6	2	6	8	-33	=СУММ(Знаки)			
4	0,591	0,665	12	13	3	16	-11	Коэффициент Кендалла				
5	0,902	0,567	20	11	4	10	0	-0,17	=2*I3/(I1*(I1-1))			
6	0,853	0,69	19	15	5	14	-7	Критериальная статистика				
7	0,667	0,808	15	18	6	20	-14	-1,21	=3*I5*КОРЕНЬ((I1*(I1-1)/(2*(I1-5))))			
8	0,305	0,199	7	4	7	4	7	Критическое значение				
9	0,151	0,666	5	14	8	8	2	1,96	=НОРМСТОБР(1-K2/2)			
10	0,714	0,537	18	9	9	12	-3	Гипотеза	принимается			
11	0,425	0,455	8	8	10	19	-10	=ЕСЛИ(ABS(D7)<D9;"принимается";"отклоняется")				
12	0,713	0,104	17	2	11	5	3					
13	0,188	0,911	6	20	12	13	-4					
14	0,433	0,61	9	12	13	1	7					
15	0,109	0,78	3	16	14	7	2					
16	0,447	0,814	10	19	15	18	-5					

Рис. 13.2. Критерий независимости на основе рангового коэффициента корреляции Кендалла

Поскольку некоторые действия выполняются вручную, для проверки гипотезы о независимости для новой выборки их придется повторить снова (это преобразование формул, вычисляющих ранги, в значения, и выполнение сортировки). Если данный критерий используется часто, то можно написать простые макросы, которые будут автоматизировать эти действия.

13.1.5. Критерий независимости для многомерных выборок

Этот критерий основан на коэффициенте согласованности (конкордации), описанном в разделе 3.3.2, и применяется для проверки гипотезы о независимости для нескольких (больше двух) случайных величин¹.

Статистическая модель. Пусть наблюдается m -мерная случайная величина $Z = (X_1, X_2, \dots, X_m)$. В результате имеем выборку объемом n $(x_{11}, x_{21}, \dots, x_{m1}), (x_{12}, x_{22}, \dots, x_{m2}), \dots, (x_{1n}, x_{2n}, \dots, x_{mn})$.

Гипотезы

H_0 : коэффициент согласованности $W = 0$;

H_1 : коэффициент согласованности $W \neq 0$.

Задан уровень значимости α .

Вычисления

1. Каждому выборочному значению $(x_{1i}, x_{2i}, \dots, x_{mi})$ присваиваются ранги $(r_{1i}, r_{2i}, \dots, r_{mi})$. Ранги r_{ji} присваиваются значениям x_{ji} независимо путем построения отдельных вариационных рядов для реализации каждого компонента X_j так же, как при вычислении коэффициентов Спирмена и Кендалла. Если есть совпадающие выборочные значения, то им присваиваются одинаковые ранги, равные среднему рангов, которые были бы им присвоены при отсутствии равенства значений.

2. Вычисляется коэффициент согласованности

$$3. W = \frac{12}{m^2(n^3 - n)} \sum_{i=1}^n \left(\sum_{j=1}^m r_{ji} - \frac{m(n+1)}{2} \right)^2.$$

4. Вычисляется критериальная статистика $T = m(n-1)W$.

Построение критической области. При условии истинности нулевой гипотезы коэффициент W имеет специальное распределение, статистика T асимптотически имеет распределение χ^2 с $(n-1)$ степенью свободы.

Определяется критическое значение t как квантиль порядка $1 - \alpha$ распределения χ^2 с $(n-1)$ степенью свободы. Гипотеза H_0 принимается, если выполняется неравенство $T \leq t$, иначе гипотеза H_0 отклоняется.

Комментарии

1. Этот критерий приближенный и применяется для выборок объемом не менее 10. Для малых выборок в качестве критериальной статистики берется коэффициент W , а критическое значение определяется из таблицы специального распределения этого коэффициента.

2. Для проверки независимости двух случайных величин рекомендуется применять критерии на основе ранговых коэффициентов корреляции Спирмена или Кендалла.

Практическая реализация

На рис. 13.3 показан рабочий лист Excel, реализующий данный критерий. В столбцах А, В и С записаны выборочные значения (выборочные значения

¹ Попутно отметим, что для многомерных выборок средство Excel Корреляция может вычислить корреляционную матрицу, а средство Коварияция — ковариационную матрицу. Эти средства описаны в главе 5.

независимы и имеют равномерные распределения на интервале $[0, 1]$). В столбцах D, E и F подсчитаны ранги с помощью функции РАНГ для значений каждого столбца в отдельности (так же, как в предыдущих критериях). Далее в ячейку G2 введена формула

$$=(\text{СУММ}(D2:F2)-\$I\$3*(\$I\$1+1)/2)^2$$

Она реализует для $i = 1$ часть формулы $\left(\sum_{j=1}^n r_{ji} - \frac{m(n+1)}{2} \right)^2$ вычисления коэффициента согласованности. Эта формула затем скопирована вниз до конца интервала G2:G21. Остальные формулы данного рабочего листа очевидны.

G2 = (СУММ(D2:F2)-\$I\$3*(\$I\$1+1)/2)^2														
	A	B	C	D	E	F	G	H	I	J	K	L		
1	X	Y	Z	Ранги		Сумма	Объем	20					Уровень значимости	0,05
2	0,76	0,66	0,02	15	14	3	0,25	Количество переменных						
3	0,23	0,87	0,01	5	19	2	30,25						3	
4	0,6	0,49	0,31	13	10	10	2,25	Общая сумма						
5	0,78	0,76	0,21	17	17	7	90,25						1767	=СУММ(Сумма)
6	0,2	0,43	0,14	4	9	4	210,25	Козффициент согласованности						
7	0,79	0,73	0,58	18	15	13	210,25						0,295238	=12*15/(13*13*11*((1*11-1)))
8	0,19	0,76	0,25	3	16	9	12,25	Критериальная статистика						
9	0,64	0,3	0,43	14	6	11	0,25						16,82857	=13*(11-1)*17
10	0,83	0,41	0,8	19	8	15	110,25	Критическое значение						
11	0,42	0,54	0,79	10	12	14	20,25						30,14351	=ХИ2ОБР(L1;11-1)
12	0,38	0,07	0,84	9	2	18	6,25	Гипотеза					принимается	
13	0,31	0,3	0,82	7	7	17	0,25						=ЕСЛИ(I9<I11;"принимается";"отклоняется")	
14	0,97	0,24	0,15	20	4	5	6,25							
15	0,76	0,61	0,97	16	13	20	306,25							
16	0,17	0,03	0,18	2	1	6	506,25							

Рис. 13.3. Критерий независимости для многомерных выборок

13.1.6. Критерий независимости на основе таблиц сопряженности

Этот критерий, иногда называемый критерием независимости χ^2 , разработан для определения независимости классификационных случайных величин (см. раздел 3.3.3). Однако его можно применять к случайным величинам других типов. В частности, он хорошо подходит для определения независимости количественных случайных величин, имеющих дискретные распределения на конечном множестве значений. Мы покажем этот критерий на примере именно дискретных случайных величин.

Статистическая модель. Пусть имеется выборка $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, которая является реализацией двумерной дискретной случайной величины $Z = (X, Y)$, где случайная величина X принимает конечное число значений A_1, A_2, \dots, A_r , а случайная величина Y — также конечное число значений B_1, B_2, \dots, B_s .

Гипотезы

H_0 : случайные величины X и Y независимы;

H_1 : случайные величины X и Y зависимы.

Задан уровень значимости α .

1. По выборке составляется таблица сопряженности следующего вида.

	A_1	A_2	...	A_s	Всего
B_1	v_{11}	v_{12}	...	v_{1s}	$n_{1\cdot} = \sum_{i=1}^s v_{1i}$
B_2	v_{21}	v_{22}	...	v_{2s}	$n_{2\cdot} = \sum_{i=1}^s v_{2i}$
...
B_r	v_{r1}	v_{r2}	...	v_{rs}	$n_{r\cdot} = \sum_{i=1}^s v_{ri}$
Всего	$n_{\cdot 1} = \sum_{i=1}^r v_{i1}$	$n_{\cdot 2} = \sum_{i=1}^r v_{i2}$...	$n_{\cdot s} = \sum_{i=1}^r v_{is}$	$n = \sum_{i=1}^r n_{i\cdot} = \sum_{i=1}^r n_{\cdot i}$

Здесь v_{ij} — количество выборочных значений (x_i, y_j) , имеющих значения A_j и B_i .

2. Вычисляется критериальная статистика

$$T = n \sum_{i=1}^r \sum_{j=1}^s \frac{(v_{ij} - n_{i\cdot} n_{\cdot j})^2}{n_{i\cdot} n_{\cdot j}} = n \left(\sum_{i=1}^r \sum_{j=1}^s \frac{v_{ij}^2}{n_{i\cdot} n_{\cdot j}} - 1 \right).$$

Построение критической области. При условии истинности нулевой гипотезы статистика T приближенно имеет распределение χ^2 со степенью свободы $(r-1)(s-1)$.

Определяется критическое значение t как квантиль порядка $1 - \alpha$ распределения χ^2 со степенью свободы $(r-1)(s-1)$. Гипотеза H_0 принимается, если выполняется неравенство $T \leq t$, иначе гипотеза H_0 отклоняется.

Комментарий

1. Это асимптотический критерий. Поэтому необходимо, чтобы объем выборки был не менее 20.
2. Существует определенная проблема, связанная с количеством наблюдений в каждой ячейке таблицы сопряженности. Можно встретить рекомендации объединять ячейки с малым количеством наблюдений. В общем случае здесь надо придерживаться тех же правил, что и в критерии χ^2 (см. раздел 9.2.1).

Практическая реализация

На рис. 13.4 показан рабочий лист Excel, реализующий данный критерий. В качестве тестовой выборки взята двумерная выборка, компоненты которой независимы и имеют распределение Пуассона с параметром $\lambda = 2$. Выборка получена с помощью средства Генерация случайных чисел. Поскольку в данном случае случайные величины принимают неотрицательные целочисленные значения, для определения количества различных значений в выборках X и Y достаточно

найти максимальные значения, имеющиеся в этих выборках. Эти значения с помощью функции МАКС получены в ячейках D1 и D2. В ячейке D3 с помощью функции СЧЁТ вычислен объем выборки.

	D5		{=СУММ((X=\$C5)*(Y=D\$4))}										
	A	B	C	D	E	F	G	H	I	J	K	L	
1	X	Y	Максимум x	7	Уровень значимости								
2	0	3	Максимум y	6	0.05								
3	1	0	Объем	30	Таблица сопряженности								
4	1	1		0	1	2	3	4	5	6	Всего		
5	1	3	0	1	0	0	2	0	0	0	3		
6	4	3	1	1	2	1	1	1	0	1	7		
7	2	0	2	2	4	2	1	1	1	0	11		
8	2	0	3	1	0	0	0	1	0	1	3		
9	0	0	4	0	1	2	1	0	0	0	4		
10	4	2	5	0	0	0	0	0	1	0	1		
11	3	4	6	0	0	0	0	0	0	0	0.01		
12	1	6	7	0	0	0	0	1	0	0	1		
13	1	4	Всего	5	7	5	5	4	2	2	30	30.01	
14	1	1	Сумма	2.3998	{=СУММ((D5:J12)^2)/(K5:K12)*(D13:J13))}								
15	4	2	Статистика	41.994	{=(D14-1)*D3}								
16	7	4	Критическое значение		{=ЕСЛИ(D15<D17,"принимается","отклоняется")}								
17	3	0		58.124	Гипотеза								
18	2	5		{=ХИ2ОБР(F2,D1*D2)}									

Рис. 13.4. Критерий независимости на основе таблицы сопряженности

Далее создается таблица сопряженности. В диапазоне D4:J4 записаны значения, которые принимает переменная Y, а в диапазоне C5:K12 — переменная X. Для вычисления значений таблицы сопряженности в ячейку D5 вводится формула массива $\{=СУММ((X=\$C5)*(Y=D\$4))\}$, которая затем копируется во все остальные ячейки таблицы сопряженности. Эта формула в ячейке D5 подсчитывает количество одновременных совпадений значений в диапазоне X со значением ячейки C5 и значений в диапазоне Y со значением ячейки D4. При наличии таких совпадений часть формулы $(X=\$C5)*(Y=D\$4)$ продуцирует единицу, в противном случае — нуль.

В диапазоне K5:K12 подсчитываются суммы значений по строкам таблицы сопряженности, а в диапазоне D13:J13 — по столбцам. Во избежание возможного деления на нуль в формуле ячейки D14 эти суммы вычисляются с использованием функции ЕСЛИ, которая записывает в ячейки значение 0,01, если сумма равна нулю. Например, в ячейке K5 записана формула

$$=ЕСЛИ(СУММ(D5:J5)=0;0,01;СУММ(D5:J5))$$

Аналогичные формулы используются в других ячейках, вычисляющих суммы по строкам и столбцам таблицы сопряженности. Значение 0,01 взято произвольно, оно никак не влияет на последующие вычисления и просто показывает, что в данной строке или столбце сумма равна нулю. В ячейке K13 для контроля вычисляется сумма диапазона D13:J13, а в ячейке L13 — сумма диапазона K5:K12. Целая часть значения этих сумм должна равняться объему выборки.

В ячейке D14 вычисляется часть формулы $\sum_{i=1}^n \sum_{j=1}^n \frac{v_{ij}}{n_i \cdot n_j}$ вычисления критерияльной статистики. Формула массива в ячейке D14 имеет вид

$$\{=СУММ(((D5:J12)^2)/((K5:K12)*(D13:J13))))\}$$

Здесь в полной мере реализуются возможности формул массивов — без использования формулы массива пришлось бы строить промежуточную таблицу (подобную таблице сопряженности), чтобы выполнить вычисления $\frac{v_{ij}^2}{n_i \cdot n_j}$ для каждой ячейки

таблицы сопряженности. После вычисления этой суммы нахождение критерияльной статистики (ячейка D15) и критического значения (ячейка D17) не представляет особых трудностей. Формулы для их вычислений показаны на рис. 13.4.

Отметим, что все формулы на этом рабочем листе “живые” и автоматически пересчитываются при изменении выборочных значений.

13.2. Оценивание коэффициента корреляции

Если с помощью критериев независимости установлен факт зависимости между случайными величинами, то далее возникает вопрос оценки степени этой зависимости. В качестве меры статистической зависимости случайных величин обычно выступает коэффициент корреляции. Поэтому вопрос об оценке степени статистической зависимости можно переформулировать в вопрос о точности значения вычисленного выборочного коэффициента корреляции. Ответ на последний вопрос дают доверительные интервалы и критерии проверки гипотез о значении коэффициента корреляции. В этом разделе рассмотрены методы построения доверительных интервалов для коэффициентов корреляции, а в следующем — методы сравнения выборочных коэффициентов корреляции.

13.2.1. Доверительные интервалы для коэффициента корреляции

Если коэффициент корреляции отличен от нуля, то точное распределение выборочного коэффициента корреляции даже в случае нормального распределения зависимых случайных величин является весьма сложным и неприменимым для практического использования. В этой ситуации только применение z -преобразования Фишера $z = \frac{1}{2} \ln \frac{1+r}{1-r}$ (см. раздел 3.3.1) предоставляет возможность построить приближенные доверительные интервалы для неизвестного коэффициента корреляции.

Статистическая модель. Выборочные значения $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ являются реализацией случайной величины $Z = (X, Y)$, имеющей произвольное двумерное распределение с конечными моментами второго порядка и с коэффициентом корреляции ρ . Объем выборки — не менее 20.

Доверительный интервал для коэффициента корреляции ρ строится следующим образом.

1. Задается доверительный уровень α .

2. Вычисляются выборочные средние $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ и $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

3. Вычисляется точечная оценка коэффициента корреляции

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}.$$

4. Из уравнения $\alpha = 2\Phi(k) - 1$, где Φ — функция распределения стандартного нормального закона, вычисляется значение k : $k = \Phi^{-1}\left(\frac{1+\alpha}{2}\right)$, Φ^{-1} — функция, обратная к функции распределения стандартного нормального закона.

5. Вычисляются величины $z_1 = \frac{1}{2} \ln \frac{1+r}{1-r} - \frac{k}{\sqrt{n-3}}$ и $z_2 = \frac{1}{2} \ln \frac{1+r}{1-r} + \frac{k}{\sqrt{n-3}}$.

6. Вычисляется доверительный интервал (r_1, r_2) , где $r_1 = \frac{e^{2z_1} - 1}{e^{2z_1} + 1}$ и $r_2 = \frac{e^{2z_2} - 1}{e^{2z_2} + 1}$.

Комментарии

1. Еще раз подчеркнем, что это приближенный метод, который дает удовлетворительные результаты для достаточно больших выборок (объемом более 20 значений).

2. Для нормальных совокупностей можно использовать метод, описанный в следующем разделе.

Практическая реализация

На рис. 13.5 показан рабочий лист Excel, на котором приведены все формулы, необходимые для вычисления доверительного интервала. Тестовая выборка имеет совместное нормальное распределение с коэффициентом корреляции 0,5. Выборка построена методом, описанным в разделе 7.5.1. Отметим, что для вычисления в ячейках D7 и F7 границ доверительного интервала по известным значениям z_1 и z_2 используется функция TANH, вычисляющая значения гиперболического тангенса — формулы $r_1 = \frac{e^{2z_1} - 1}{e^{2z_1} + 1}$ и $r_2 = \frac{e^{2z_2} - 1}{e^{2z_2} + 1}$ являются формулами гиперболического тангенса. Такие же вычисления выполняет функция ФИШЕРОБР. Для вычисления преобразования Фишера в Excel предусмотрена специальная функция ФИШЕР, которая использована в формуле ячейки D3. О функциях ФИШЕР и ФИШЕРОБР речь идет в разделе 4.10.5.

гиперболического тангенса. Такие же вычисления выполняет функция ФИШЕРОБР. Для вычисления преобразования Фишера в Excel предусмотрена специальная функция ФИШЕР, которая использована в формуле ячейки D3. О функциях ФИШЕР и ФИШЕРОБР речь идет в разделе 4.10.5.

гиперболического тангенса. Такие же вычисления выполняет функция ФИШЕРОБР. Для вычисления преобразования Фишера в Excel предусмотрена специальная функция ФИШЕР, которая использована в формуле ячейки D3. О функциях ФИШЕР и ФИШЕРОБР речь идет в разделе 4.10.5.

13.2.2. Доверительные интервалы для коэффициента корреляции нормальной совокупности

Статистическая модель. Выборочные значения $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ являются реализацией случайной величины $Z = (X, Y)$, имеющей двумерное нормальное распределение с коэффициентом корреляции ρ . Объем выборки — не менее 20.

	A	B	C	D	E	F	G	H
	X	Y	Объем	30	Доверительный уровень			
2	-1,22719	0,293671	r =	0,587914		0,95		
3	-0,52777	-0,6682	Z =	0,674473	Коэффициент k			
4	-2,07131	-1,72014			1,959963	=НОРМСТОБР((1+F2)/2)		
5	-0,17293	-0,11686	z1	0,297278	z2	1,051668	=D3+F4/КОРЕНЬ(D1-3)	
6	0,07308	-0,05764	Доверительный интервал					
7	0,793681	1,173913	r1	0,288819	r2	0,782454	=TANH(F5)	
8	0,141907	0,313723	=TANH(D5)					
9	1,013197	-0,01488						
10	-0,80214	0,143778	Формулы:					
11	0,746446	0,020679	в ячейке D1 =СЧЕТ(X)					
12	2,079083	0,008004	в ячейке D2 =КОРРЕЛ(X;Y)					
13	0,073672	-0,38575	в ячейке D3 =ФИШЕР(D2)					
14	-1,28082	-0,81158	в ячейке D5 =D3-F4/КОРЕНЬ(D1-3)					
15	0,197218	-1,26596						
	-0,16779	0,354865						

Рис. 13.5. Построение доверительного интервала для коэффициента корреляции

Доверительный интервал для коэффициента корреляции ρ строится следующим образом.

1. Задается доверительный уровень α .

2. Вычисляются выборочные средние $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ и $\bar{y} = \frac{1}{n} \sum_{i=1}^n y_i$.

3. Вычисляется точечная оценка коэффициента корреляции

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}$$

4. Из уравнения $\alpha = 2\Phi(k) - 1$, где Φ — функция распределения стандартного нормального закона, определяется значение k : $k = \Phi^{-1}\left(\frac{1+\alpha}{2}\right)$, Φ^{-1} — функция, обратная к функции распределения стандартного нормального закона.

5. Вычисляется доверительный интервал

$$\left(r + \frac{r(1-r^2)}{2n} - k \frac{1-r^2}{\sqrt{n}}, r + \frac{r(1-r^2)}{2n} + k \frac{1-r^2}{\sqrt{n}} \right).$$

Комментарии

1. Это асимптотический метод, поэтому объем выборки должен быть больше 20. Метод построен на основе того факта, что в случае нормальной совокупности выборочный коэффициент корреляции асимптотически имеет нормальное распределение с математическим ожиданием ρ и дисперсией

$\left(\frac{1-\rho^2}{\sqrt{n}}\right)^2$. Поправка $\frac{r(1-r^2)}{2n}$ введена из-за смещения выборочного коэффициента корреляции r относительно истинного коэффициента корреляции ρ .

2. Метод плохо работает, если ρ близко к ± 1 . В этом случае объем выборки должен быть очень большим, чтобы распределение r с приемлемой точностью аппроксимировалось нормальным распределением.
3. Если ρ близко к нулю, для построения доверительного интервала можно применить распределение Стюдента, как это сделано в критерии независимости для нормальных совокупностей (см. раздел 13.1.2). Но такой доверительный интервал также будет приближенным.

Практическая реализация

На рис. 13.6 показан рабочий лист Excel, на котором приведены все формулы, необходимые для построения доверительного интервала.

D1		=СЧЕТ(X)					
	A	B	C	E	F	G	H
	X	Y	Объем	30	Доверительный уровень		
2	0.307841	0.964899	r=	0,511411	0,95		
3	-1,61556	-1,24569	=КОРРЕЛ(X;Y)	Коэффициент k			
4	0,394347	0,455268			1,959963	=НОРМСТОБР((1+F2)/2)	
5	0,359116	-1,54766	Доверительный интервал				
6	-1,51117	-0,4825	r1	0,253456	r2	0,781954	
7	-0,71308	-0,81218	=D2+D2*(1-D2*D2)/(2*D1)+F4*(1-D2*D2)/КОРЕНЬ(D1)				
8	0,304692	-0,15631	=D2+D2*(1-D2*D2)/(2*D1)-F4*(1-D2*D2)/КОРЕНЬ(D1)				
9	-0,58185	1,000847					
10	-0,3608	0,768231					
11	-0,90482	-1,06931					
12	-0,50718	-1,2386					

Рис. 13.6. Построение доверительного интервала для коэффициента корреляции нормальной совокупности

13.3. Критерии проверки гипотез о значениях коэффициента корреляции

Сделаем общий комментарий ко всем описанным в этом разделе критериям проверки гипотез о значениях коэффициента корреляции. Все эти критерии построены с использованием z -преобразования Фишера. Поэтому все они являются асимптотическими и требуют, чтобы выборки были достаточно большого объема. Для проверки гипотезы о равенстве нулю коэффициента корреляции используются критерии проверки независимости, которые описаны в разделе 13.1.

13.3.1. Критерий проверки значения коэффициента корреляции

Статистическая модель. Выборочные значения $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ являются реализацией случайной величины $Z = (X, Y)$, имеющей произвольное

двумерное распределение с коэффициентом корреляции ρ . Объем выборки — не менее 20.

Гипотезы

а) Равенство	б) Неравенство	в) Неравенство
$H_0: \rho = \rho_0$	$H_0: \rho \leq \rho_0$	$H_0: \rho \geq \rho_0$
$H_1: \rho \neq \rho_0$	$H_1: \rho > \rho_0$	$H_1: \rho < \rho_0$

Здесь ρ_0 — заданное число. Задан уровень значимости α .

Вычисления

1. Вычисляется точечная оценка коэффициента корреляции

$$r = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2 \cdot \sum_{i=1}^n (y_i - \bar{y})^2}}, \text{ где } \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \text{ и } \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

2. Вычисляются величины $z = \frac{1}{2} \ln \frac{1+r}{1-r}$ и $z_0 = \frac{1}{2} \ln \frac{1+\rho_0}{1-\rho_0}$.

3. Вычисляется критериальная статистика $T = \frac{z - z_0}{\sqrt{n-3}}$.

Построение критической области. При условии истинности нулевых гипотез статистика T асимптотически имеет стандартное нормальное распределение.

Случай а). Определяются критические значения t как квантиль порядка $1 - \alpha/2$ стандартного нормального распределения. Гипотеза H_0 принимается, если $|T| \leq t$. В противном случае гипотеза H_0 отклоняется.

Случай б). Определяется критическое значение t_1 как квантиль порядка $1 - \alpha$ стандартного нормального распределения. Гипотеза H_0 принимается, если $T \leq t_1$.

Случай в). Определяется критическое значение t_2 как квантиль порядка α стандартного нормального распределения. Гипотеза H_0 принимается, если $t_2 \leq T$.

Практическая реализация в Excel этого критерия не вызывает затруднений.

13.3.2. Критерий проверки равенства двух коэффициентов корреляции

Статистическая модель. Заданы две выборки, сделанные из двумерных совокупностей с коэффициентами корреляции ρ_1 и ρ_2 соответственно. Объем первой выборки равен n_1 , объем второй — n_2 .

Гипотезы

а) Равенство	б) Неравенство	в) Неравенство
$H_0: \rho_1 = \rho_2$	$H_0: \rho_1 \leq \rho_2$	$H_0: \rho_1 \geq \rho_2$
$H_1: \rho_1 \neq \rho_2$	$H_1: \rho_1 > \rho_2$	$H_1: \rho_1 < \rho_2$

Задан уровень значимости α .

1. Вычисляются точечные оценки коэффициентов корреляции первой выборки r_1 и второй выборки r_2 по стандартным формулам.

2. Вычисляются величины $z_1 = \frac{1}{2} \ln \frac{1+r_1}{1-r_1}$, $z_2 = \frac{1}{2} \ln \frac{1+r_2}{1-r_2}$ и $S = \sqrt{\frac{1}{n_1-3} + \frac{1}{n_2-3}}$.

3. Вычисляется критериальная статистика $T = \frac{z_1 - z_2}{S}$.

Построение критической области. При условии истинности нулевых гипотез статистика T асимптотически имеет стандартное нормальное распределение.

Случай а). Определяются критические значения t как квантиль порядка $1 - \alpha/2$ стандартного нормального распределения. Гипотеза H_0 принимается, если $|T| \leq t$. В противном случае гипотеза H_0 отклоняется.

Случай б). Определяется критическое значение t_1 как квантиль порядка $1 - \alpha$ стандартного нормального распределения. Гипотеза H_0 принимается, если $T \leq t_1$.

Случай в). Определяется критическое значение t_2 как квантиль порядка α стандартного нормального распределения. Гипотеза H_0 принимается, если $t_2 \leq T$.

Практическая реализация в Excel этого критерия показана на рис. 13.7. Здесь же приведены все необходимые для вычислений формулы.

F1 =СЧЕТ(X1)										
	A	B	C	D	E	F	G	H	I	J
1	X1	Y1	X2	Y2	Объем 1	30	Уровень значимости			
2	-0,68	-0,464	0,8407	0,8534	r1 =	0,1252	0,05			
3	0,3686	0,3661	-1,492	-0,816	Объем 2	25	Статистика			
4	0,7176	0,6841	-0,582	-1,172	r2 =	0,6911	-2,52177=(F5-F6)/F7			
5	0,4339	-0,607	0,2333	0,4725	z1 =	0,1259	Критическое значение 1			
6	0,9665	1,1006	0,1444	0,3047	z2 =	0,8501	1,958963=НОРМСТОБР(1-H2/2)			
7	-1,108	1,6672	0,483	-1,493	S =	0,2872	Критическое значение 2			
8	0,1808	-1,039	1,4506	-0,044			1,644853=НОРМСТОБР(1-H2)			
9	-0,006	-0,986	1,2802	1,0468	Гипотеза а	отклоняется	=ЕСЛИ(ABS(H4)<H8,"принимается","отклоняется")			
10	0,428	1,5374	0,5224	-0,128	Гипотеза б	принимается	=ЕСЛИ(H4<H8,"принимается","отклоняется")			
11	0,5506	-0,586	1,4143	0,9575	Гипотеза в	отклоняется	=ЕСЛИ(H4>H8,"принимается","отклоняется")			
12	1,3023	-0,061	1,5534	1,3667						
13	-0,256	1,2429	-0,265	-1,617	Формулы					
14	0,3258	0,535	-0,124	-1,158	в ячейке F2	=КОРРЕЛ(X1;Y1)				
15	1,1732	0,4146	0,0317	0,2867	в ячейке F4	=КОРРЕЛ(X2;Y2)				
16	-0,89	-0,04	1,3394	0,4701	в ячейке F5	=ФИШЕР(F2)				
17	1,4874	1,8506	-1,226	-0,121	в ячейке F6	=ФИШЕР(F4)				
18	1,2219	0,2148	-1,028	-1,535	в ячейке F7	=КОРЕНЬ(1/(F1-3)+1/(F3-3))				
	-0,392	-0,922	0,9852	0,4967						

Рис. 13.7. Критерий проверки равенства двух коэффициентов корреляции

13.3.3. Критерий проверки равенства нескольких коэффициентов корреляции

Статистическая модель. Заданы k выборок, сделанных из двумерных совокупностей с коэффициентами корреляции $\rho_1, \rho_2, \dots, \rho_k$ соответственно. Объем первой выборки равен n_1 , объем второй — n_2 , ..., k -й выборки — n_k .

Гипотезы

$H_0: \rho_1 = \rho_2 = \dots = \rho_k$;

H_1 : нулевая гипотеза неверна.

Вычисления

1. Вычисляются точечные оценки коэффициентов корреляции всех выборок r_1, r_2, \dots, r_k по стандартным формулам.

2. Вычисляются величины $z_1 = \frac{1}{2} \ln \frac{1+r_1}{1-r_1}$, $z_2 = \frac{1}{2} \ln \frac{1+r_2}{1-r_2}$, ..., $z_k = \frac{1}{2} \ln \frac{1+r_k}{1-r_k}$.

3. Вычисляется критериальная статистика $T = \sum_{i=1}^k (n_i - 3) z_i^2 - \frac{\left(\sum_{i=1}^k (n_i - 3) z_i \right)^2}{\sum_{i=1}^k (n_i - 3)}$.

Построение критической области. При условии истинности нулевых гипотез статистика T асимптотически имеет распределение χ^2 с $(k - 1)$ степенью свободы.

Определяются критические значения t как квантиль порядка $1 - \alpha$ распределения χ^2 с $(k - 1)$ степенью свободы. Гипотеза H_0 принимается, если $T \leq t$. В противном случае гипотеза H_0 отклоняется.

Комментарий. В случае $k = 2$ критерий эквивалентен критерию а) из предыдущего раздела, при этом значение критериальной статистики данного критерия равняется квадрату статистики критерия из предыдущего раздела.

Практическая реализация в Excel этого критерия не вызывает затруднений.

Сравнение зависимых выборок

В этой главе рассмотрены методы сравнения параметров распределений зависимых выборок. Если методами из главы 13 установлен факт зависимости выборочных значений, то методы сравнения параметров распределений, описанные в главе 12, применять нельзя. Для зависимых выборок существуют специальные методы. Им и посвящена данная глава. В первом разделе рассмотрены методы построения доверительных интервалов для разностей математических ожиданий, во втором — критерии проверки гипотез о значениях математических ожиданий и в третьем — методы дисперсионного анализа.

14.1. Доверительные интервалы для разности математических ожиданий нормальных совокупностей

Точные доверительные интервалы для разности математических ожиданий зависимых выборок известны только для случая нормально распределенных генеральных совокупностей. Для произвольных распределений можно применить непараметрические критерии сравнения математических ожиданий, которые хотя и не строят доверительные интервалы, но позволяют проверить гипотезы о равенстве или неравенстве этих разностей некоторым заданным значениям.

14.1.1. Доверительный интервал для разности математических ожиданий

Статистическая модель. Выборочные значения $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ являются реализацией случайной величины $Z = (X, Y)$, имеющей двумерное нормальное распределение. Случайная величина X имеет математическое ожидание μ_1 , случайная величина Y — математическое ожидание μ_2 .

Доверительный интервал для разности $\Delta\mu = \mu_1 - \mu_2$ строится следующим образом.

1. Задается доверительный уровень α .
2. Вычисляются разности $x_1 - y_1, x_2 - y_2, \dots, x_n - y_n$.
3. Вычисляется среднее этих разностей $\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$ и их выборочная дисперсия $s_d^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2$.

$$\text{для } s_d^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2.$$

4. Определяется коэффициент k — квантиль порядка $(1 + \alpha)/2$ распределения Стьюдента с $(n - 1)$ степенью свободы.
5. Вычисляется доверительный интервал $\left(\bar{d} - k \frac{S_d}{\sqrt{n}}, \bar{d} + k \frac{S_d}{\sqrt{n}} \right)$.

Комментарии

1. Доверительный интервал строится на том основании, что разность нормальных случайных величин (даже зависимых) также будет иметь нормальное распределение.
2. Метод устойчив при умеренных отклонениях от нормальности.
3. Метод неприменим для независимых наблюдений.

Практическая реализация

На рис. 14.1 показан рабочий лист, на котором реализован данный способ построения доверительного интервала. В столбцах A и B содержатся выборочные значения, имеющие совместное нормальное распределение с нулевыми математическими ожиданиями и коэффициентом корреляции 0,5 (способы создания таких выборок показаны в разделе 7.5.1). Таким образом, здесь разность математических ожиданий равна нулю. Диапазон ячеек, содержащий значения первой выборки, назван X, а второй выборки — Y. Обращаем внимание на формулы массивов в ячейках D3 и D5, с помощью которых вычисляются среднее разностей и выборочная дисперсия разностей. Применение этих формул позволяет избежать промежуточных вычислений самих разностей.

D3		{=CPЗНАЧ(X-Y)}					
A	B	C	D	E	F	G	H
1	X	Y	Объем	30	Доверительный уровень		
2	-1,2811	-1,1044	Среднее разностей		0,95		
	-1,2804	-0,4715	-0,0583726	Коэффициент k			
4	-0,4295	0,27884	Дисперсия разностей		2,363849	=СТЮДРАСПОБР((1-F2)/2;D1-1)	
			0,76457342	{=ДИСП(X-Y)}			
			Доверительный интервал				
7	1,36609	1,57159	-0,4357438	0,318999	=D3+F4*КОРЕНЬ(D5/D1)		
8	-1,6951	-0,3377	=D3-F4*КОРЕНЬ(D5/D1)				
9	-0,1739	0,3598					
10	0,53269	1,14435					
11	-1,0299	-0,8232					
12	-0,1418	0,33955					

Рис. 14.1. Построение доверительного интервала для разностей математических ожиданий

14.1.2. Доверительный интервал для математических ожиданий нескольких совокупностей

Статистическая модель. Пусть наблюдается m -мерная нормально распределенная случайная величина $Z = (X_1, X_2, \dots, X_m)$. В результате имеем выборку объемом n $(x_{11}, x_{21}, \dots, x_{m1}), (x_{12}, x_{22}, \dots, x_{m2}), \dots, (x_{1n}, x_{2n}, \dots, x_{mn})$. Обозначим через $\mu_1, \mu_2, \dots, \mu_m$ неизвестные математические ожидания случайных величин X_1, X_2, \dots, X_m .

Доверительный интервал строится для линейной комбинации математических ожиданий, т.е. для величины $\bar{L} = c_1\mu_1 + c_2\mu_2 + \dots + c_m\mu_m$, где c_1, c_2, \dots, c_m — заданные числа, сумма которых равна нулю.

Доверительный интервал строится следующим образом.

1. Задается доверительный уровень α .

2. Вычисляются m средних вида $\bar{x}_{1.} = \frac{1}{n} \sum_{i=1}^n x_{1i}, \bar{x}_{2.} = \frac{1}{n} \sum_{i=1}^n x_{2i}, \dots, \bar{x}_{m.} = \frac{1}{n} \sum_{i=1}^n x_{mi}.$

3. Вычисляется величина $\bar{L} = c_1\bar{x}_{1.} + c_2\bar{x}_{2.} + \dots + c_m\bar{x}_{m.}.$

4. Вычисляются n средних вида $\bar{x}_{.1} = \frac{1}{m} \sum_{i=1}^m x_{i1}, \bar{x}_{.2} = \frac{1}{m} \sum_{i=1}^m x_{i2}, \dots, \bar{x}_{.n} = \frac{1}{m} \sum_{i=1}^m x_{in}.$

5. Вычисляется общее среднее $\bar{x} = \frac{1}{mn} \sum_{i=1}^m \sum_{j=1}^n x_{ij}.$

6. Вычисляется сумма квадратов $S = \sum_{i=1}^m \sum_{j=1}^n (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2$ и дисперсия

$$s^2 = \frac{S}{(m-1)(n-1)}.$$

7. Определяется величина t как квантиль порядка $1 - \alpha$ F -распределения со степенями свободы 1 и $(n-1)(m-1)$.

8. Вычисляется величина $T = \sqrt{\frac{t(c_1^2 + c_2^2 + \dots + c_m^2)s^2}{n}}.$

9. Вычисляется доверительный интервал $(\bar{L} - T, \bar{L} + T).$

Комментарии

1. В основе данного метода построения доверительного интервала лежит двухфакторный дисперсионный анализ (см. разделы 3.5.3 и 14.3).

2. Метод устойчив при умеренных отклонениях от нормальности и при небольших отклонениях от условия равенства дисперсий.

3. Этим методом можно строить доверительные интервалы для попарных разностей математических ожиданий. Например, если положить $c_1 = 1, c_2 = -1$, а все остальные c_i положить равными нулю, то будет построен доверительный интервал для разности математических ожиданий первых двух выборок.

Практическая реализация

На рис. 14.2 показан рабочий лист Excel, на котором реализован данный метод построения доверительного интервала для линейной комбинации математических ожиданий. В столбцах A:D записаны исходные данные — выборка из четырех зависимых компонентов, имеющих нормальное распределение. Диапазоны ячеек, содержащие выборочные значения этих компонентов, названы соответственно X1, X2, X3 и X4.

F2 F (=СУММКВ(A2:D2-\$G\$7:\$J\$7-E2+\$H\$9))											
	A	B	C	D	E	F	G	H	I	J	K
1	X1	X2	X3	X4	Средние	Квадраты	Объем (n)	30	Доверительный уровень		
	-0,714	-0,468	-0,713	-2,017	-0,97602	1,544639	m	4	0,95		
3	-0,005	0,262	0,497	1,426	0,54254	1,033317	c1	c2	c3	c4	
4	-0,868	0,089	0,235	-1,373	-0,42923	1,547993	1	1	-1	-1	
5	-0,773	-1,043	-0,463	-1,867	-1,03853	1,178571	Средние по столбцам				
6	-0,959	-1,066	-0,884	-0,452	-0,84023	0,220546	X1	X2	X3	X4	
7	-0,222	-0,416	0,125	0,371	-0,0353	0,356792	-0,2403893	-0,04668	-0,05585	-0,0815	
8	-0,982	0,065	-0,411	0,471	-0,2143	0,841725	Общее среднее				
9	0,798	0,954	1,472	1,192	1,10379	0,185729	-0,1086	=CP3НАЧ(X1,X2,X3,X4)			
10	-0,251	0,563	-0,299	1,915	0,48189	3,017563	Дисперсия	0,662624	=(СУММ(Квадраты)/((H1-1)*(H2-1)))		
11	0,341	0,092	-0,235	-1,036	-0,20855	1,257324	L	-0,13971	=СУММПРОИЗВ(G4:J4,G7:J7)		
12	-1,273	-0,158	-0,465	-1,454	-0,83732	0,996549	Коэффициент k				
13	-0,439	-0,456	-0,117	-0,84	-0,48788	0,371053	3,950589	=FРАСПОБР(1-J2;1;(H1-1)*(H2-1))			
14	-1,509	-0,419	0,203	0,766	-0,23992	2,489848	Доверительный интервал				
15	0,942	0,283	-1,857	-0,054	-0,12134	4,038912	-0,7304968	0,451085	=H11+КОРЕНЬ(H13*H10*		
16	0,734	-0,485	0,046	-0,032	0,06569	1,034345	СУММКВ(G4:J4)/H1)				
17	-0,953	-1,378	-0,28	-1,522	-1,0327	0,880105	=H11-КОРЕНЬ(H13*H10*СУММКВ(G4:J4)/H1)				
18	-0,082	1,045	-0,607	-1,563	-0,30156	3,553993					
19	-0,136	0,417	-0,942	-0,678	-0,25918	0,820785					

Рис. 14.2. Построение доверительного интервала для линейной комбинации математических ожиданий

К сожалению, для реализации данного метода нельзя обойтись без некоторых промежуточных вычислений. В столбце E вычислены средние по строкам (формула =CP3НАЧ(A2:D2) в ячейке E2, которая затем скопирована вниз), в ячейках G7:J7 — средние по столбцам (формула =CP3НАЧ(X1) в ячейке G7; аналогичные формулы содержатся в остальных ячейках этого диапазона), общее среднее — в ячейке H9.

В столбце F вычисляются квадраты $\sum_{i=1}^m (x_{ij} - \bar{x}_{i.} - \bar{x}_{.j} + \bar{x})^2$ (часть формулы для вычисления S; см. приведенный выше п. 6 последовательности построения доверительного интервала). Для этого в ячейке F2 введена формула массива

$$(\text{=СУММКВ(A2:D2-G7:J7-E2+H9)}),$$

которая затем скопирована вниз. Сама величина S вычисляется как часть формулы в ячейке H10, в которой находится значение дисперсии s^2 . Формулы для других вычислений представлены на рабочем листе.

На рис. 14.2 показан доверительный интервал для разности $\mu_1 + \mu_2 - \mu_3 - \mu_4$. Изменяя коэффициенты c_i (диапазон G4:J4), можно сразу получить доверительные интервалы для других комбинаций математических ожиданий. Например, на рис. 14.3 показан доверительный интервал для разности $\mu_1 - \mu_4$.

14.2. Критерии проверки гипотез о равенстве математических ожиданий

Как показано в предыдущем разделе, точные доверительные интервалы для разностей математических ожиданий зависимых выборок можно построить только при обременительном предположении о нормальном распределении генеральной совокупности. Точные критерии проверки гипотез о равенстве математических

ожиданий также строятся на основе такого же предположения. Однако существуют непараметрические критерии, которые не требуют предположений о нормальности распределений. В этом отношении критерии предпочтительнее доверительных интервалов для сравнения математических ожиданий зависимых наблюдений.

E2		=CP3HAY(A2:D2)													
	A	B	C	D	E	F	G	H	I	J	K				
1	X1	X2	X3	X4	Средние	Квадраты	Объем (n)	30	Доверительный уровень						
2	-0,714	-0,468	-0,713	-2,017	-0,97802	1,544639	m	4	0,95						
3	-0,005	0,262	0,487	1,428	0,54254	1,033317	c1	c2	c3	c4					
4	-0,688	0,089	0,235	-1,373	-0,42923	1,547993	1	0	0	-1					
5	-0,773	-1,043	-0,463	-1,887	-1,03653	1,178571	Средние по столбцам								
6	-0,958	-1,066	-0,884	-0,452	-0,84023	0,220546	X1	X2	X3	X4					
7	-0,222	-0,416	0,125	0,371	-0,0353	0,356792	-0,2403893	-0,04668	-0,06585	-0,0815					
8	-0,982	0,085	-0,411	0,471	-0,2143	0,941725	Общее среднее								
9	0,796	0,854	1,472	1,192	1,10379	0,185728	-0,1086								
10	-0,251	0,563	-0,299	1,915	0,48189	3,017563	Дисперсия	0,862624							
11	0,341	0,092	-0,235	-1,036	-0,20855	1,257324	L	-0,15889							
12	-1,273	-0,158	-0,465	-1,454	-0,83732	0,996549	Коэффициент k								
13	-0,438	-0,456	-0,117	-0,94	-0,48788	0,371053	3,950599								
14	-1,509	-0,419	0,203	0,766	-0,23992	2,468848	Доверительный интервал								
15	0,942	0,283	-1,657	-0,054	-0,12134	4,036912	-0,5766407	0,258864							
16	0,734	-0,485	0,046	-0,032	0,06589	1,034345									
17	-0,953	-1,376	-0,28	-1,522	-1,0327	0,980105									

Рис. 14.3. Доверительный интервал для разности $\mu_1 - \mu_2$

14.2.1. Парный критерий Стьюдента

Статистическая модель. Выборочные значения $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ являются реализацией случайной величины $Z = (X, Y)$, имеющей двумерное нормальное распределение. Случайная величина X имеет математическое ожидание μ_1 , случайная величина $Y - \mu_2$.

Гипотезы

а) Равенство

$$H_0: \mu_1 - \mu_2 = m$$

$$H_1: \mu_1 - \mu_2 \neq m$$

б) Неравенство

$$H_0: \mu_1 - \mu_2 \leq m$$

$$H_1: \mu_1 - \mu_2 > m$$

в) Неравенство

$$H_0: \mu_1 - \mu_2 \geq m$$

$$H_1: \mu_1 - \mu_2 < m$$

Здесь m — заданное число. Задан уровень значимости α .

Вычисления

1. Вычисляются разности $x_1 - y_1, x_2 - y_2, \dots, x_n - y_n$.

2. Вычисляется среднее этих разностей $\bar{d} = \frac{1}{n} \sum_{i=1}^n d_i$ и их выборочная диспер-

$$сия $S_d^2 = \frac{1}{n-1} \sum_{i=1}^n (d_i - \bar{d})^2$.$$

3. Вычисляется критерияльная статистика $T = \frac{\sqrt{n}(\bar{d} - m)}{S_d}$.

Построение критической области. При условии истинности нулевых гипотез статистика T имеет распределение Стьюдента с $(n - 1)$ степенью свободы.

Случай а). Определяются критические значения t как квантили порядка $1 - \alpha/2$ распределения Стьюдента с $(n - 1)$ степенью свободы. Гипотеза H_0 принимается, если $|T| \leq t$. В противном случае гипотеза H_0 отклоняется.

Случай б). Определяется критическое значение t_1 как квантиль порядка $1 - \alpha$ распределения Стьюдента с $(n - 1)$ степенью свободы. Гипотеза H_0 принимается, если $T \leq t_1$.

Случай в). Определяется критическое значение t_2 как квантиль порядка α распределения Стьюдента с $(n - 1)$ степенью свободы. Гипотеза H_0 принимается, если $t_2 \leq T$.

Комментарии

1. Для проверки гипотезы о том, что $\mu_1 = \mu_2$, в гипотезе а) следует положить $m = 0$.
2. Критерий неприменим для независимых выборок.
3. Критерий не чувствителен к умеренным отклонениям от нормальности.
4. При значительных отклонениях от нормальности следует применять непараметрический критерий знаков или критерий Уилкоксона (см. следующие разделы).

Практическая реализация критерия в Excel не представляет трудностей и во многом совпадает с вычислением доверительного интервала из раздела 14.1.1. Кроме того, в Excel реализация этого критерия осуществляется с помощью средства Парный двухвыборочный t-тест для средних, описание которого дано в разделе 5.9. Там же приведен пример реализации критерия.

14.2.2. Непараметрический критерий знаков

Этот критерий применяется для сравнения местоположения распределений компонентов случайной величины $Z = (X, Y)$. Мерой различия в местоположении распределений служит медиана случайной величины $X - Y$. Поскольку для большинства распределений медиана и математическое ожидание близки, особенно если распределения симметричны или хотя бы одномодальны, то в случае, когда различие незначимо (т.е. медиана близка к нулю), можно считать, что математические ожидания случайных величин X и Y также различаются незначимо.

Статистическая модель. Выборочные значения $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ являются реализацией случайной величины $Z = (X, Y)$, имеющей двумерное распределение. Рассматриваются разности $x_1 - y_1, x_2 - y_2, \dots, x_n - y_n$.

Гипотезы

H_0 : медиана разностей равна нулю;

H_1 : медиана разностей не равна нулю.

Задан уровень значимости α .

Вычисления

1. Подсчитывается количество N положительных разностей $x_1 - y_1, x_2 - y_2, \dots, x_n - y_n$.

2. Для малых выборок число N берется в качестве критериальной статистики. Для больших выборок критериальная статистика вычисляется по формуле

$$T = \frac{2N - n}{\sqrt{n}}.$$

Построение критической области. При условии истинности нулевой гипотезы статистика N имеет биномиальное распределение с параметрами n и $p = 0,5$, статистика T асимптотически имеет стандартное нормальное распределение.

Для малых выборок в качестве критических значений t_n и t_n берутся соответственно квантили порядка $\alpha/2$ и $1 - \alpha/2$ биномиального распределения с параметрами n и $p = 0,5$. Гипотеза H_0 принимается, если $t_n \leq N \leq t_n$. В противном случае нулевая гипотеза отвергается.

Для больших выборок в качестве критического значения t берется квантиль порядка $1 - \alpha/2$ стандартного нормального распределения. Гипотеза H_0 принимается, если $|T| \leq t$. В противном случае нулевая гипотеза отвергается.

Комментарии

1. Если есть разности $x_i - y_i$, равные нулю, то за каждую нулевую разность к значению N надо прибавить 0,5.
2. В данном критерии большой считается выборка объемом более 20 значений.
3. С помощью данного критерия можно проверять другие гипотезы. Например, H_0 : медиана разностей равна δ (заданное число), H_1 : медиана разностей не равна δ . Для проверки таких гипотез из каждой разности $x_i - y_i$ необходимо вычесть δ . Остальные вычисления остаются без изменений.
4. Если нет оснований отклонять предположение о нормальности генеральной совокупности, то следует применять более мощный парный критерий Стьюдента (см. предыдущий раздел). Непараметрический критерий Уилкоксона, описанный ниже, также более мощный, но он предполагает выполнения более сильного условия, чем критерий знаков, а именно — симметричность распределений.

Практическая реализация

На рис. 14.4 показан рабочий лист Excel, реализующий данный критерий для больших выборок. Все формулы, необходимые для вычислений, приведены на этом листе.

D2		=СУММ(ЕСЛИ(X>Y;1;ЕСЛИ(X=Y;0,5;0)))					
	A	B	C	D	E	F	G
1	X	Y	Объем	30	Уровень значимости		
2	-0,1329	0,10588	N	14		0,05	
3	0,47155	1,92998	T	-0,365148	Критическое значение		
4	2,88264	-0,4347		= (2*D2-D1)/КОРЕНЬ(D1)	1,959963	=НОРМСТОБР(1-F2/2)	
5	1,0195	-0,5814					
6	1,70679	0,85863		Гипотеза	принимается		
7	0,73766	-0,5916			=ЕСЛИ(ABS(D3)<F4;"принимается";"отклоняется")		
8	0,9446	-0,4928					
9	0,68594	0,90458					
10	0,13805	-0,2332					
11	1,21609	1,09769					
12	-0,0192	-0,4501					

Рис. 14.4. Непараметрический критерий знаков

14.2.3. Непараметрический критерий Уилкоксона

Этот критерий применяется для сравнения математических ожиданий μ_1 и μ_2 компонентов случайной величины $Z = (X, Y)$. Однако проверяемая с помощью данного критерия гипотеза H_0 состоит в том, что распределение разностей $x_1 - y_1, x_2 - y_2, \dots, x_n - y_n$ симметрично относительно нуля (тогда математическое ожидание разностей равно нулю и, следовательно, $\mu_1 = \mu_2$). Если же эта гипотеза отклоняется, то вывод, что $\mu_1 \neq \mu_2$, можно сделать лишь тогда, когда выполняется предположение о том, что распределения компонентов X и Y имеют один и тот же вид (различны только значения математического ожидания) либо эти распределения симметричны относительно математического ожидания. На практике, однако, допустимы умеренные отклонения от выполнения указанных требований, так как критерий незначительно чувствителен к ним.

Статистическая модель. Выборочные значения $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$ являются реализацией случайной величины $Z = (X, Y)$, имеющей двумерное распределение. Рассматриваются разности $x_1 - y_1, x_2 - y_2, \dots, x_n - y_n$.

Гипотезы

H_0 : распределение разностей симметрично относительно нуля;

H_1 : нулевая гипотеза неверна.

Задан уровень значимости α .

Вычисления

1. По совокупности модулей разностей $d_1 = |x_1 - y_1|, d_2 = |x_2 - y_2|, \dots, d_n = |x_n - y_n|$ строится вариационный ряд $d_{(1)} \leq d_{(2)} \leq \dots \leq d_{(n)}$, по которому определяются ранги величин d_i . Равным по абсолютной величине разностям приписываются средние ранги. Нулевые разности игнорируются, при этом значение n уменьшается на количество нулевых разностей.
2. Подсчитывается сумма N рангов, которым соответствуют положительные разности $x_i - y_i$.

3. Вычисляется критериальная статистика $T = \frac{N - n(n+1)/4}{\sqrt{n(n+1)(2n+1)/24}}$.

Построение критической области. При условии истинности нулевой гипотезы величина N имеет специальное распределение Уилкоксона, статистика T асимптотически имеет стандартное нормальное распределение.

В качестве критического значения t берется квантиль порядка $1 - \alpha/2$ стандартного нормального распределения. Гипотеза H_0 принимается, если $|T| \leq t$. В противном случае нулевая гипотеза отвергается.

Комментарии

1. Критерий является асимптотическим и применяется для больших выборок (объемом более 20). Для малых выборок критериальной статистикой является величина N , а критическое значение определяется по таблице специального распределения Уилкоксона [14].
2. С помощью данного критерия можно проверять другие гипотезы. Например H_0 : распределение разностей симметрично относительно δ , где δ — заданное число (это равносильно гипотезе, что $\mu_1 - \mu_2 = \delta$); H_1 : нулевая гипотеза неверна.

неверна. Для проверки таких гипотез из каждой разности $x_i - y_i$ необходимо вычесть 8. Остальные вычисления остаются без изменений.

3. Если нет оснований отклонять предположение о нормальности генеральной совокупности, то следует применять более мощный парный критерий Стьюдента (см. раздел 14.2.1).

Практическая реализация

Реализация данного критерия в Excel показана на рис. 14.5, на котором приведены все формулы, необходимые для вычисления критерия. Отметим, что здесь невозможно обойтись без промежуточных вычислений: в столбце С вычисляются модули попарных разностей выборочных значений, в столбце D — ранги этих модулей. В ячейке E3 вычисляется количество совпадающих парных значений, поскольку при совпадении объем выборки уменьшается на число таких совпадений.

C2			f. (=ABS(X-Y))								
	A	B	C	D	E	F	G	H	I	J	K
1	X	Y	Модули	Ранг	Объем	"30	Уровень значимости				
2	0,053	0,067	0,01464	2	N =	180	I 0,05				
3	0,192	1,346	1,15425	25	n0 =	0	Критическое значение				
"4	-1,3	0,128	1,42588	28:	n =	30	1.959963 =НОРМСТОБР(1-F2/2)				
5	-0,49	0,532	1,02389	24!	T =	-1,07984					
6	1,016	1,235	0,21899	7!	Гипотеза		принимается				
>	-0,14	-0,33	0,19026	6!	=ЕСЛИ(ABS(D3)<F4;"npMHHMaеTCH";						отклоняется")
8	-0,65	-0,2	0,44221	"ТГГ""							
9"	1,015	1,591	0,57582	21	Формулы						
T6'	-1,31'	-1,61	0,2986	11; в ячейке F1 =СЧЁТ(X)							
11	0,872	0,456	0,41641	15! в ячейке F2 {=СУММ(Е'KX>Y;PaHr;""))}							
"h \	-0,64	-0,73	0,08386	4! в ячейке P3{=СУММ(ЕСЛИ(X=Y;1;""))} I							
~13	0,987"	-1,21	2,2005	30; в ячейке F4 =F1-F3							!
14	-0,97	-0,63	0,33463	14' в ячейке F5 =(F2-F4*(F-4+1)/4)/КОРЕНЬ(F4*(F4+1) j*(2*(F4+1)/24)...							
"15 •	0,704	0,164	0,54075	20 в диапазоне Модули {=ABS(X-Y)} j							
~16j	-0,22	0,078	0,29307	10 в диапазоне "анг" {=РАНГ(Модули;Модули; 1)}							
"TFI	2,215	2,176	0,03956з! ""•							

Рис. 14.5. Не параметрический критерий Уилкоксона

14.3. Дисперсионный анализ для зависимых выборок

Основные статистические предположения, на которых строится дисперсионный анализ (см. раздел 3.5.1), заключаются в том, что ошибки наблюдений независимы и имеют нормальное распределение с нулевым математическим ожиданием и одинаковыми дисперсиями. Исследования влияния нарушений основных предположений на выводы дисперсионного анализа [24, гл. 10] показывают, что дисперсионный анализ наиболее чувствителен к нарушениям предположений о нормальности распределений и равенстве дисперсий и наименее чувствителен к нарушениям предположения о независимости наблюдений¹. Поэтому на практике дисперсионный анализ часто применяется к зависимым выборкам.

¹ Подчеркнем, что здесь речь идет не о модели со случайными факторами (см. раздел 3.5.1), а о модели с постоянными факторами, но случайными ошибками наблюдений. В модели со случайными факторами применяется схема вычислений, отличная от схемы вычислений в модели с постоянными факторами.

14.3.1. Двухфакторный дисперсионный анализ

Статистическая модель. Имеется двумерная выборка, состоящая из выборочных значений x_{ij} ; индекс i соответствует уровню β_i фактора β , индекс j соответствует уровню γ_j фактора γ . Пусть фактор β имеет r уровней, а фактор γ — t уровней; выборка имеет размерность $r \times t$. Таким образом, каждое выборочное значение x_{ij} можно представить в виде

$$x_{ij} = \mu + \beta_i + \gamma_j + \varepsilon_{ij},$$

где μ — константа (общее среднее), ε_{ij} — случайные величины, имеющие нормальное распределение с нулевым математическим ожиданием и одинаковыми дисперсиями. Все величины ε_{ij} независимы.

Гипотезы

а) Равенство значений уровней фактора β б) Равенство значений уровней фактора γ

$H_0: \beta_1 = \beta_2 = \dots = \beta_r;$

$H_0: \gamma_1 = \gamma_2 = \dots = \gamma_t;$

H_1 : не все значения уровней равны.

H_1 : не все значения уровней равны.

Задан уровень значимости α .

Вычисления

1. Выборку удобно представлять в виде таблицы, по значениям которой вычисляются средние по строкам и столбцам и общее среднее.

	γ_1	γ_2	...	γ_t	Средние
β_1	x_{11}	x_{12}	...	x_{1t}	$\bar{x}_{1\cdot} = \frac{1}{t} \sum_{j=1}^t x_{1j}$
β_2	x_{21}	x_{22}	...	x_{2t}	$\bar{x}_{2\cdot} = \frac{1}{t} \sum_{j=1}^t x_{2j}$
...
β_r	x_{r1}	x_{r2}	...	x_{rt}	$\bar{x}_{r\cdot} = \frac{1}{t} \sum_{j=1}^t x_{rj}$
Средние	$\bar{x}_{\cdot 1} = \frac{1}{r} \sum_{i=1}^r x_{i1}$	$\bar{x}_{\cdot 2} = \frac{1}{r} \sum_{i=1}^r x_{i2}$...	$\bar{x}_{\cdot t} = \frac{1}{r} \sum_{i=1}^r x_{it}$	$\bar{x} = \frac{1}{rt} \sum_{i=1}^r \sum_{j=1}^t x_{ij}$

2. Вычисляются компоненты дисперсионной таблицы.

Источник вариации (компоненты дисперсии)	Сумма квадратов	Число степеней свободы	Дисперсия
Вариация между средними по строкам (различия между уровнями фактора β)	$SS_1 = t \sum_{i=1}^r (\bar{x}_{i\cdot} - \bar{x})^2$	$r - 1$	$s_1^2 = \frac{SS_1}{r - 1}$

Источник вариации (компоненты дисперсии)	Сумма квадратов	Число степеней свободы	Дисперсия
Вариация между средними по столб- цам (различия между уровнями фактора γ)	$SS_2 = r \sum_{i=1}^t (\bar{x}_{.i} - \bar{x})^2$	$t - 1$	$s_2^2 = \frac{SS_2}{t - 1}$
Остаточная вари- ация (различия внутри выборки)	$SS_3 = \sum_{i=1}^r \sum_{j=1}^t (x_{ij} - \bar{x}_{.i} - \bar{x}_{.j} + \bar{x})^2$	$(r - 1)(t - 1)$	$s_3^2 = \frac{SS_3}{(r - 1)(t - 1)}$
Полная (общая) ва- риация	$SS = \sum_{i=1}^r \sum_{j=1}^t (x_{ij} - \bar{x})^2$	$rt - 1$	$s^2 = \frac{SS}{rt - 1}$

3. Вычисляются критериальные статистики $T_\beta = \frac{s_1^2}{s_3^2}$ и $T_\gamma = \frac{s_2^2}{s_3^2}$.

Построение критической области. При условии истинности нулевых гипотез статистика T_β имеет F -распределение со степенями свободы $(r - 1)$ и $(r - 1)(t - 1)$, статистика T_γ имеет F -распределение со степенями свободы $(t - 1)$ и $(r - 1)(t - 1)$.

Случай а). Определяется критическое значение $t_{кр}$ как квантиль порядка $1 - \alpha$ F -распределения со степенями свободы $(r - 1)$ и $(r - 1)(t - 1)$. Если выполняется неравенство $T_\beta < t_{кр}$, то нулевая гипотеза принимается, в противном случае — отклоняется.

Случай б). Определяется критическое значение $t_{кр}$ как квантиль порядка $1 - \alpha$ F -распределения со степенями свободы $(t - 1)$ и $(r - 1)(t - 1)$. Если выполняется неравенство $T_\gamma < t_{кр}$, то нулевая гипотеза принимается, в противном случае — отклоняется.

Комментарии

1. Критерий устойчив при умеренных отклонениях от требования нормальности и равенства дисперсий.
2. Для выборок, для которых не выполняются требования нормальности и равенства дисперсий, необходимо применять непараметрический критерий Фридмана из следующего раздела.
3. Могут проверяться другие гипотезы, например $H_0: \beta_1 - \beta_2 = a$, $\beta_2 = \beta_3 = \dots = \beta_r$; H_1 : нулевая гипотеза не верна (a — заданное число). В этом случае сначала число a вычитается из выборочных значений x_{ij} ($i = 1, 2, \dots, t$), а затем выполняются все вычисления критерия без изменений. Аналогичные гипотезы можно проверять относительно значений уровней фактора γ .
4. Если нулевая гипотеза отклоняется, значит, не все значения уровней фактора одинаковы. Для того чтобы определить, какие значения уровней фактора отличаются от других, следует применить метод множественных сравнений Шеффе из раздела 14.3.3.

Практическая реализация

В Excel данный критерий реализует описанное в разделе 5.13 средство Двухфакторный дисперсионный анализ без повторений из пакета анализа. Там же приведен пример его применения. Здесь покажем выполнение критерия без средства Двухфакторный дисперсионный анализ без повторений. Его использование имеет тот недостаток, что при изменении выборочных значений или для другого уровня значимости его необходимо применять заново. Рабочий лист, показанный на рис. 14.6, лишен этого недостатка — любые изменения в выборке приводят к автоматическому пересчету критерия. Например, на рис. 14.7 изменен первый столбец выборочных значений (“волеуриарисиски” введены единицы²) — рабочий лист ауматическис пересчисан и получен новый результат. Здесь верно отвергася гипотеза о равенстве значений уровней фактора γ .

F8	α* = CP3HA4(B3:E7)								
A	B	C	D	E		G	H	I	J
1	Выборка					r =	5		
2	Бета 1	Бета 2	Бета 3	Бета 4	Средние	t =	4		
3	Гамма 1	7,025997	39,0232	44,031	35,4835	31,390915	Уровень значимости		
4	Гамма 2	22,44087	6,31421	39,578	15,5309	20,966002		0,05	
5	Гамма 3	8,02589	24,9412	20,6053	9,954184	15,881585	Тв =	0,8892	=D11/D13
6	Гамма 4	8,934192	14,1277	4,12927	41,70589	17,224269	Тγ =	0,2899	=D12/D13
7	Гамма 5	39,6024	44,138	17,1554	9,302168	27,549494	Критические значения		
8	Средние	17,20583	25,7089	25,0998	22,39533	22,602453	тб =	3,2582	=FРАСПОБР(H4;C11;C13)
9	Дисперсионная таблица						тγ =	3,4903	
10	Σ квадратов	df	Дисперсия					=FРАСПОБР(H4;C12;C13)	
11	Фактор β	713,9327	4	178,483	=B11/C11	Гипотеза β	принимается		
12	Фактор γ	225,2646	3	75,0882	=B12/C12	Гипотеза γ	принимается		
13	Остаток	3107,788	12	258,982	=B13/C13	=ЕСЛИ(H6<H9;"принимается";"отвергается")			
14	Полная	4046,985	19	212,999	=B14/C14	=ЕСЛИ(H5<H8;"принимается";"отвергается")			
15	=C11*C12 =H1*H2-1								
16	Ячейка B11 (=H2*СУММКВ(F3:F7;F8))								
17	Ячейка B12 (=H1*СУММКВ(B8:E8;F8))								
18	Ячейка B13 =B14-B12-B11								
19	Ячейка B14 (=СУММКВ(B3:E7;F8))								
20									

Рис. 14.6. Двухфакторный дисперсионный анализ

Все формулы, необходимые для вычисления критерия, показаны на рис. 14.6. В столбце df дисперсионной таблицы вычисляются степени свободы соответствующих сумм квадратов. Обращаем внимание, что остаточная сумма квадратов (ячейка B13) вычисляется, как разность между полной суммой квадратов (ячейка B14) и суммой квадратов, вычисленных для факторов (ячейки B11 и B12).

14.3.2. Двухфакторный дисперсионный анализ Фридмана

Если предположения, на которых основан двухфакторный дисперсионный анализ (см. раздел 3.5.3), не выполняются, используется непараметрический критерий Фридмана. Но необходимо отметить, что обычный дисперсионный анализ более мощный, чем данный критерий. Поэтому критерий Фридмана применяется только тогда, когда есть веские основания отвергнуть статистическую модель обычного дисперсионного анализа

Эти данные не удовлетворяют условиям применимости рассматриваемого критерия.

С11		$f_n = H-1$							
	A	B	C	D	E	F	G	H	I
1	Выборка						$r =$	5	
2		Бета 1	Бета 2	Бета 3	Бета 4	Средние	$t =$	4	
3	Гамма 1	1	39,0232	44,031	35,4835	29,88442	Уровень значимости		
4	Гамма 2	1	6,31421	39,57802	15,5309	15,60578			
5	Гамма 3	1	24,9412	20,60527	9,954184	14,12516	$T\beta =$	0,0533	
6	Гамма 4	1	14,1277	4,12927	41,70589	15,24072	$T\gamma =$	3,5274	
7	Гамма 5	1	44,138	17,15544	9,302168	17,8989	Критические значения		
8	Средние	1	25,7089	25,0998	22,39533	18,551	$t\beta =$	3,2592	
9	Дисперсионная таблица						$t\gamma =$	3,4903	
10	Σ квадратов		df	Дисперсия					
11	Фактор β	672,367	4	168,0918		Гипотеза β	принимается		
12	Фактор γ	2084,69	3	694,8969		Гипотеза γ	отвергается		
13	Остаток	2364,02	12	197,0014					
14	Полная	5121,08	19	269,5303					
15									

Рис. 14.7. Двухфакторный дисперсионный анализ для новых данных

Статистическая модель. Имеется двумерная выборка, состоящая из выборочных значений x_{ij} ; индекс i соответствует уровню β_i фактора β , индекс j соответствует уровню γ_j фактора γ . Пусть фактор β имеет r уровней, а фактор γ — t уровней; выборка имеет размерность $rx \cdot t$. Таким образом, каждое выборочное значение x_{ij} можно представить в виде

$$x_{ij} = \mu + \beta_i + \gamma_j + \varepsilon_{ij},$$

где μ — константа (общее среднее), ε_{ij} — случайные величины, имеющие одинаковые распределения с одинаковыми дисперсиями (нормальность распределения не предполагается). Все величины ε_{ij} независимы.

Гипотезы

а) Равенство значений уровней фактора β б) Равенство значений уровней фактора γ

$H_0: \beta_1 = \beta_2 = \dots = \beta_r;$

$H_0: \gamma_1 = \gamma_2 = \dots = \gamma_t;$

$H_1:$ не все значения уровней равны.

$H_1:$ не все значения уровней равны.

Задан уровень значимости α .

Вычисления

Выборку удобно представлять в виде таблицы.

	γ_1	γ_2	...	γ_t
β_1	x_{11}	x_{12}	...	x_{1t}
β_2	x_{21}	x_{22}	...	x_{2t}
...
β_r	x_{r1}	x_{r2}	...	x_{rt}

1. Вычисления для проверки гипотезы а).

а) В каждом столбце таблицы по отдельности вычисляются ранги выборочных значений.

б) Вычисляется сумма рангов выборочных значений каждой строки R_1, R_2, \dots, R_r .

в) Вычисляется величина $S_r = \sum_{i=1}^r R_i^2 - \frac{1}{r} \left(\sum_{i=1}^r R_i \right)^2$.

г) Вычисляется критериальная статистика $T_r = \frac{12S_r}{tr(r+1)}$.

2. Вычисления для проверки гипотезы б).

а) В каждой строке таблицы по отдельности вычисляются ранги выборочных значений.

б) Вычисляется сумма рангов выборочных значений каждого столбца $\bar{R}_1, \bar{R}_2, \dots, \bar{R}_t$.

в) Вычисляется величина $S_t = \sum_{i=1}^t \bar{R}_i^2 - \frac{1}{t} \left(\sum_{i=1}^t \bar{R}_i \right)^2$.

г) Вычисляется критериальная статистика $T_t = \frac{12S_t}{rt(t+1)}$.

Построение критической области. При условии истинности нулевых гипотез величины S_r и S_t имеют специальное распределение Фридмана, статистики T_r и T_t асимптотически имеют распределение χ^2 со степенями свободы $(r-1)$ и $(t-1)$ соответственно.

Для малых выборок критические значения определяются по таблицам распределения Фридмана [14]. В случае больших выборок ($r \geq 5, t \geq 5$) для проверки гипотезы а) критическое значение $t_{кр}$ определяется как квантиль порядка $1 - \alpha$ распределения χ^2 со степенью свободы $(r-1)$. Если $T_r \leq t_{кр}$, то нулевая гипотеза принимается. В противном случае — отвергается. Для проверки гипотезы б) критическое значение $t_{кр}$ находится как квантиль порядка $1 - \alpha$ распределения χ^2 со степенью свободы $(t-1)$. Если $T_t \leq t_{кр}$, то нулевая гипотеза принимается. В противном случае — отвергается.

Комментарии

1. Если в какой-либо строке или столбце имеются одинаковые значения, то им присваиваются средние ранги.

2. Критерий мало чувствителен к умеренным отклонениям от требования одинаковой распределенности величин ϵ_{ij} .

Практическая реализация

На рис. 14.8 показан рабочий лист Excel, реализующий данный критерий. В качестве исходных данных использована выборка из примера предыдущего раздела. К сожалению, в данном случае не удалось обойтись без промежуточных вычислений. В диапазоне G3:J7 вычисляются ранги выборочных значений по столбцам. Для этого сначала выделяется диапазон G3:G7 (как показано на рис. 14.8) и вводится формула массива $\{=РАНГ(B3:B7;B3:B7;1)\}$, которая затем копируется в ячейки всего диапазона G3:J7. Аналогично в диапазоне B9:E13 вычисляются ранги по строкам — здесь сначала в диапазон B9:E9 вводится формула массива $\{=РАНГ(B3:E3;B3:E3;1)\}$, которая затем копируется в ячейки всего

диапазона В9:Е13. В диапазонах К3:К7 и В14:Е14 вычисляются суммы соответствующих рангов. Остальные формулы приведены на рис. 14.8. Если критерий выполняется часто, то, чтобы освободить данный рабочий лист, диапазоны Г3:К7 и В9:Е14, содержащие ранги выборочных значений и их суммы, можно перенести в отдаленную область листа. Вычисления от этого не пострадают.

G3 (=РАНГ(B3:B7;B3:B7;1))												
	A	B	C	D	E	F	G	H	I	J	K	L
1	Выборка										r =	5
2		Бета 1	Бета 2	Бета 3	Бета 4	Бета 1	Бета 2	Бета 3	Бета 4	Сумма	t =	4
3	Гамма 1	7,026	39,023	44,031	35,484	1	4	5	4	14	Уровень значимости	
4	Гамма 2	22,441	6,3142	39,578	15,531	4	1	4	3	12	0,05	
5	Гамма 3	8,0257	24,941	20,605	9,9542	2	3	3	2	10		
6	Гамма 4	8,9342	14,128	4,1293	41,708	3	2	1	5	11	=СУММ(G6:J8)	
7	Гамма 5	39,502	44,139	17,155	9,3022	5	5	2	1	13	=СУММ(G7:J7)	
8	=СУММКВ(К3:К7)-((СУММ(К3:К7))^2)/M1											
9	Гамма 1	1	3	4	2	St =	10	Ty =	1,2	=12*N9/(M1*M2*(M2+1))		
10	Гамма 2	3	1	4	2	Критическое значение			7,815 =Х12ОБР(М4;М2-1)			
11	Гамма 3	1	4	3	2	Гипотеза β			принимается =ЕСЛИ(Е15<D16,"принимается","от			
12	Гамма 4	2	3	1	4	Гипотеза γ			принимается =ЕСЛИ(J8<K10,"принимается","отв			
13	Гамма 5	3	4	2	1							
14	Сумма	10	15	14	11	=СУММ(Е9:Е13)						
15	Sr =		17	Tβ =	1,7	=12*C15/(M2*M1*(M1+1))						
16	Критическое значение: 9,4977 =Х12ОБР(М4;М1-1)											
17	=СУММКВ(В14:Е14)-((СУММ(В14:Е14))^2)/M2											
18												

Рис. 14.8. Непараметрический критерий Фридмана

На рис. 14.9 показан тот же рабочий лист с измененными исходными данными: значения в первом столбце выборки заменены единичными значениями. Как и следовало ожидать, критерий отклонил гипотезу о равенстве значений уровней фактора β, но только при уровне значимости 0,1³.

	A	B	C	D	E	F	G	H	I	J	K	L	
1	Выборка											r = 5	
2	Бета 1 Бета 2 Бета 3 Бета 4				Бета 1 Бета 2 Бета 3 Бета 4				Сумма	t = 4			
3	Гамма 1	1	39,023	44,031	35,484	1	4	5	4	14	Уровень значимости		
4	Гамма 2	1	6,3142	39,578	15,531	1	1	4	3	9	0,1		
5	Гамма 3	1	24,941	20,605	9,9542	1	3	3	2	9			
6	Гамма 4	1	14,128	4,1293	41,708	1	2	1	5	9			
7	Гамма 5	1	44,138	17,155	9,3022	1	5	2	1	9			
8													
9	Гамма 1	1	3	4	2	St = 20		Ty = 2,4					
10	Гамма 2	1	2	4	3	Критическое значение				6,251			
11	Гамма 3	1	4	3	2								
12	Гамма 4	1	3	2	4	Гипотеза β отвергается							
13	Гамма 5	1	4	3	2	Гипотеза γ принимается							
14	Сумма	5	16	16	13								
15	Sr = 81		T β = 8,1										
16	Критическое значение				7,7794								
17													

Рис. 14.9. Непараметрический критерий Фридмана для новых данных

Эти данные не удовлетворяют условиям применимости рассматриваемого критерия.

14.3.3. Критерий множественных сравнений Шеффе для зависимых выборок

Двухфакторный дисперсионный анализ позволяет обнаруживать разные значения уровней факторов, однако не представляет возможности указывать, какой именно уровень выделяется в ряду остальных уровней. Для решения этой задачи нельзя выполнить серию последовательных попарных сравнений с помощью, например, парного критерия Стьюдента, поскольку в серии попарных сравнений резко возрастает групповая вероятность отклонения нулевой гипотезы в случае ее истинности. Попарные сравнения следует выполнять с помощью критерия множественных сравнений Шеффе.

Статистическая модель. Имеется двумерная выборка, состоящая из выборочных значений x_{ij} ; индекс i соответствует уровню β_i фактора β , индекс j соответствует уровню γ_j фактора γ . Пусть фактор β имеет r уровней, а фактор γ — t уровней; выборка имеет размерность $r \times t$. Таким образом, каждое выборочное значение x_{ij} можно представить в виде

$$x_{ij} = \mu + \beta_i + \gamma_j + \varepsilon_{ij},$$

где μ — константа (общее среднее), ε_{ij} — случайные величины, имеющие нормальное распределение с нулевым математическим ожиданием и одинаковыми дисперсиями. Все величины ε_{ij} независимы.

Гипотезы

$H_0: c_1\beta_1 + c_2\beta_2 + \dots + c_r\beta_r$, где c_1, c_2, \dots, c_r — заданные числа, сумма которых равна нулю;

H_1 : нулевая гипотеза неверна.

Задан уровень значимости α .

Вычисления в значительной мере повторяют вычисления двухфакторного дисперсионного анализа (см. раздел 14.3.1): сначала вычисляются средние по строкам $\bar{x}_{i\cdot}$ ($i = 1, 2, \dots, r$) и столбцам $\bar{x}_{\cdot j}$ ($j = 1, 2, \dots, t$) и общее среднее \bar{x} . Далее вычисляются компоненты дисперсионной таблицы; хотя для дальнейших вычислений необходима только остаточная дисперсия s_3^2 , ее сложно вычислить без остальных компонентов дисперсионной таблицы.

Вычисляется критерияльная статистика $T = \frac{\sum_{i=1}^r c_i \bar{x}_{i\cdot}}{(r-1)s_3^2 \sum_{i=1}^r c_i^2/t}$.

Построение критической области. При условии истинности нулевой гипотезы статистика T имеет F -распределение со степенями свободы $(r-1)$ и $(r-1)(t-1)$.

Определяется критическое значение $t_{кр}$ как квантиль порядка $1-\alpha$ F -распределения со степенями свободы $(r-1)$ и $(r-1)(t-1)$. Если выполняется неравенство $T < t_{кр}$, то нулевая гипотеза принимается. В противном случае — отклоняется.

Комментарии

1. Критерий обычно применяется для серии сравнений типа $H_0: \beta_1 - \beta_2 = 0$; $H_0: \beta_1 - \beta_2 \neq 0$.

2. Очевидно использование этого критерия для сравнения значений уровней фактора γ .

Практическая реализация

На рис. 14.10 показан рабочий лист Excel, на котором реализован данный критерий. Коэффициенты c , задаются в диапазоне G6:J6. Формулы для вычисления компонентов дисперсионной таблицы показаны на рис. 14.6.

	A	B	C	D	E	F	G	H	I	J	K
1	Выборка						$\alpha \approx$	5			
2		Бета 1	Бета 2	Бета 3	Бета 4	Средние	$t \approx$	4			
3	Гамма 1	7,026	39,02	44,03	35,484	31,39091	Уровень значимости				
4	Гамма 2	22,44	6,314	39,58	15,531	20,966		0,05			
5	Гамма 3	8,026	24,94	20,61	9,9542	15,88158	$c1$	$c2$	$c3$	$c4$	
6	Гамма 4	8,934	14,13	4,129	41,706	17,22427	1	-1	0	0	
7	Гамма 5	39,6	44,14	17,16	9,3022	27,54949	Статистика				
8	Средние	17,21	25,71	25,1	22,395	22,60245		0,1396			
9	Дисперсионная таблица						Критические значения				
10	Σ квадратов		df	Дисперсия			$t \approx$	3,2592	=FРАСПОБР(Н4;С11;С13)		
11	Фактор β	713,9	4	178,5			Гипотеза принимается				
12	Фактор γ	225,3	3	75,09			=ЕСЛИ(Н8<Н10;"принимается";"отвергается")				
13	Остаток	3108	12	259							
14	Полная	4047	19	213							
15											
16	Формула в ячейке Н8: =((СУММПРОИЗВ(В8:Е8;G6:J6))^2)/((Н1-1)*D13*СУММКВ(G6:J6)/Н2)										
17											

Рис. 14.10. Критерий множественных сравнений Шеффе

Регрессионный анализ

Регрессионный анализ выполняется в рамках модели, в которой переменные X и Y (возможно, векторозначные) связаны зависимостью $Y(X) = f(X) + \varepsilon$, где ε — случайная переменная. Это уравнение называется *уравнением регрессии*, а функция $f(X)$ — *функцией регрессии*. Относительно случайной величины ε обычно делается предположение, что она имеет нормальное распределение с нулевым математическим ожиданием. Эта модель и основные понятия регрессионного анализа описаны в разделе 3.4.

Пусть имеются исходные данные (наблюдения) $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$, где x_i и y_i могут быть векторами. Методы регрессионного анализа используются для решения следующих задач.

1. Подбор функции регрессии $f(X)$, которая наилучшим образом аппроксимирует исходные данные. Критерием наилучшего подбора обычно выступает критерий минимума суммы квадратов (раздел 3.4.1). При этом, как правило, функцию $f(X)$ выбирают таким образом, чтобы она имела вид

$$f(X) = b_0\phi_0(X) + b_1\phi_1(X) + b_2\phi_2(X) + \dots + b_m\phi_m(X),$$

где функции ϕ_i заданы. Коэффициенты b_i определяются на основе исходных данных методом наименьших квадратов (см. раздел 3.4.2). Конечно, ничто не мешает выбирать функции регрессии из другого класса функций или использовать другой метод вычисления коэффициентов b_i . Однако такие функции линейны относительно неизвестных коэффициентов b_i , что значительно облегчает вычисление значений этих коэффициентов. Кроме того, значения коэффициентов, вычисленные по методу наименьших квадратов обладают хорошими статистическими свойствами (если выполняется предположение о нормальном распределении случайной величины ε), что дает возможность строить для них доверительные интервалы и проверять гипотезы о их значимости.

2. Проверка гипотез о статистической значимости уравнения регрессии, т.е. проверка того, что выбранная функция регрессии адекватно описывает зависимость между переменными X и Y .
3. Проверка гипотез о статистической значимости коэффициентов регрессии. В частности, если все коэффициенты незначимо отличаются от нуля, можно утверждать, что между переменными X и Y нет зависимости, по крайней мере такой, какую можно представить в виде выбранной функции регрессии.
4. Построение доверительных интервалов для значений коэффициентов регрессии. Такие интервалы показывают точность найденных значений

коэффициентов. Это особенно важно, если коэффициенты имеют определенный “физический” смысл в рамках определенной интерпретации экспериментальных данных.

5. Определение значения переменной Y при тех значениях переменной X , которые отсутствуют в исходных данных. Это задача прогнозирования или восстановления значений (см. раздел 3.4.5).

Практические методы решения описанных задач приведены в следующих разделах главы. Сразу отметим, что в Excel имеется достаточно средств для решения данных задач, поэтому практически не возникает необходимости создавать собственные формулы — достаточно применить имеющиеся функции и средства.

15.1. Построение функции регрессии

Пусть имеются исходные данные (наблюдения) $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. Предполагаем, что исходя из каких-либо соображений выбран вид функции регрессии $Y(X) = b_0\varphi_0(X) + b_1\varphi_1(X) + b_2\varphi_2(X) + \dots + b_m\varphi_m(X)$, где функции φ_i известны (заданы), и нужно вычислить коэффициенты b_0, b_1, \dots, b_m . В соответствии с методом наименьших квадратов для этого необходимо решить систему нормальных уравнений (см. раздел 3.4.2)

$$\begin{cases} b_0 \sum_{i=1}^n \varphi_0^2(x_i) + b_1 \sum_{i=1}^n \varphi_0(x_i)\varphi_1(x_i) + \dots + b_m \sum_{i=1}^n \varphi_0(x_i)\varphi_m(x_i) = \sum_{i=1}^n y_i \varphi_0(x_i); \\ b_0 \sum_{i=1}^n \varphi_0(x_i)\varphi_1(x_i) + b_1 \sum_{i=1}^n \varphi_1^2(x_i) + \dots + b_m \sum_{i=1}^n \varphi_1(x_i)\varphi_m(x_i) = \sum_{i=1}^n y_i \varphi_1(x_i); \\ \dots \\ b_0 \sum_{i=1}^n \varphi_0(x_i)\varphi_m(x_i) + b_1 \sum_{i=1}^n \varphi_1(x_i)\varphi_m(x_i) + \dots + b_m \sum_{i=1}^n \varphi_m^2(x_i) = \sum_{i=1}^n y_i \varphi_m(x_i). \end{cases}$$

Для решения подобных систем сначала следует подсчитать все суммы, которые имеются в этой системе, и затем применить одно из средств Excel: использовать матричные вычисления (см. раздел 6.1.5) или средство Поиск решения (раздел 6.3.3). Попутно отметим, что с помощью средства Поиск решения можно находить параметры (коэффициенты) функций регрессии, нелинейных относительно этих параметров.

Вместе с тем Excel позволяет находить коэффициенты регрессии без построения нормальной системы уравнений. Основными средствами Excel, вычисляющими коэффициенты регрессии, являются функция ЛИНЕЙН (см. раздел 4.9.1) и средство Регрессия (раздел 5.16). Они могут вычислить коэффициенты любой функции регрессии, линейной относительно этих коэффициентов. Для этого необходимо, чтобы исходные данные имели определенную структуру, а именно, чтобы в отдельных диапазонах были заранее вычислены значения $\varphi_i(x_i)$. Примеры таких структур данных для множественной регрессии приведены при описании функции ЛИНЕЙН и средства Регрессия. Здесь покажем структуру данных для функции нелинейной регрессии $Y = X^2 + 2\ln(1 + X)$.

На рис. 15.1 представлены исходные данные: в столбце A записаны значения x_i , в столбце B — значения x_i^2 , в столбце C — значения $\ln(1 + x_i)$. Значения y_i

получены по формуле $y_i = x_i^2 + 2 \times \ln(1 + x_i) + \epsilon_i$, где ϵ_i — реализации случайной величины, имеющей стандартное нормальное распределение. Коэффициенты b_1 и b_2 функции регрессии $Y = b_1 X^2 + b_2 \ln(1 + X)$ определены с помощью функции ЛИНЕЙН (истинные значения коэффициентов $b_1 = 1$ и $b_2 = 2$). Сначала выделяется диапазон ячеек, в который будут записаны значения коэффициентов, затем вводится функция ЛИНЕЙН с необходимыми аргументами и нажимается комбинация клавиш <Ctrl+Shift+Enter>, поскольку функция вводится, как формула массива.

F3 {=ЛИНЕЙН(D2:D16;A2:C16;;1)}									
	A	B	C	D	Y = b0 + b1*X + b2*X^2 + b3*ln(1+X)				
1	X	X^2	ln(1+X)	Y					
2	1,14	1,3	0,833	1,084	b3	b2	b1	b0	
3	1,056	1,114	0,749	3,109	5,66757	0,95495	-2,7205	0,10874	
4	2,987	8,924	2,295	12,99	5,08624	0,38946	4,63022	1,08685	
5	2,95	8,703	2,272	13,58	R^2 0,96101	1,10996	#N/D	#N/D	
6	2,573	6,622	2,031	11,6	F 90,3825	11	#N/D	#N/D	
7	0,532	0,283	0,249	-0,188	334,054	13,552	#N/D	#N/D	
8	0,442	0,195	0,178	1,08	Y = b0 + b1*X^2 + b2*ln(1+X)				
9	0,003	6E-06	6E-06	0,22	b2	b1	b0		
10	0,352	0,124	0,117	-0,988	2,78784	0,85355	-0,4205	{=ЛИНЕЙН(D2:D16;B2:C16;;1)}	
11	1,92	3,686	1,545	7,244	1,32207	0,33947	0,59146		
12	0,821	0,675	0,516	3,121	R^2 0,95979	1,07925	#N/D		
13	1,209	1,462	0,901	4,668	F 143,216	12	#N/D		
14	1,92	3,685	1,544	7,126	333,629	13,9773	#N/D		
15	2,324	5,399	1,856	9,019	Y = b1*X^2 + b2*ln(1+X)				
16	0,992	0,983	0,685	0,832	b2	b1	b0		
17					2,07309	1,00104	0 {=ЛИНЕЙН(D2:D16;B2:C16;0;1)}		
18					0,84202	0,26353	#N/D		
				R^2	0,9581	1,05852	#N/D		
				F	148,617	13	#N/D		
					333,04	14,566	#N/D		

Рис. 15.1. Вычисление коэффициентов регрессии

В диапазоне F3:I7 введена формула {=ЛИНЕЙН(D2:D16;A2:C16;;1)}; здесь в качестве исходных данных использованы значения из столбцов A:C. (Обращаем внимание, что заголовки столбцов не включаются в исходные данные, иначе функция возвращает значение ошибки #ЧИСЛО!) При таких исходных данных вычисляются коэффициенты b_0 , b_1 , b_2 и b_3 функции регрессии $Y = b_0 + b_1 X + b_2 X^2 + b_3 \ln(1 + X)$. Как видно на рис. 15.1, вычисленные значения коэффициентов b_2 и b_3 весьма далеки от истинных значений. Если столбец A не включать в исходные данные, то будут вычислены коэффициенты b_0 , b_1 и b_2 функции $Y = b_0 + b_1 X^2 + b_2 \ln(1 + X)$, как это сделано в диапазоне F10:H14 с помощью формулы {=ЛИНЕЙН(D2:D16;B2:C16;;1)}. Здесь коэффициенты b_1 и b_2 также еще далеки от истинных. Если в последней формуле в качестве третьего аргумента функции ЛИНЕЙН указать 0 (это означает, что принудительно полагается $b_0 = 0$), то получатся значения коэффициентов b_1 и b_2 , весьма близкие к истинным (ячейки F17 и G17).

Хотя в последнем случае значения коэффициентов b_1 и b_2 близки к истинным, коэффициент детерминации R^2 , определяющий степень точности аппроксимации исходных данных функцией регрессии (см. раздел 3.4.3), наименьший

среди трех вычисленных (ячейки F5, F12 и F19). Это вполне объяснимо: чем больше членов в функции регрессии, тем точнее аппроксимация.

Аналогичные результаты можно получить с помощью средства Регрессия. Преимущество использования функции ЛИНЕЙН по сравнению со средством Регрессия состоит в том, что при изменении исходных данных формулы, построенные на основе функции ЛИНЕЙН, автоматически пересчитываются, в то время как средство Регрессия пришлось бы применять повторно.

Кроме функции ЛИНЕЙН и средства Регрессия, в Excel имеются и другие средства вычисления коэффициентов регрессии. Это функция ЛГРФПРИБЛ, которая вычисляет коэффициенты $b_0, m_1, m_2, \dots, m_k$ экспоненциальной регрессии вида $Y = b_0 \cdot m_1^{x_1} \cdot m_2^{x_2} \cdot \dots \cdot m_k^{x_k}$ (см. раздел 4.9.6); функция также вычисляет статистические показатели регрессии. Функции ОТРЕЗОК и НАКЛОН (см. раздел 4.9.2) вычисляют соответственно коэффициенты b и m уравнения линейной регрессии $Y = b + mX$.

В Excel функцию регрессии можно построить непосредственно на графике зависимости Y от X , построенном по экспериментальным данным. Роль функции регрессии выполняет линия тренда (см. раздел 6.2.1). Таким способом можно построить функцию регрессии только одной переменной. Этот недостаток компенсируется широким набором типов функции регрессии:

- линейная — функция регрессии вида $Y = b + mX$;
- логарифмическая — функция регрессии вида $Y = b + m \ln(X)$;
- полиномиальная — функция регрессии вида $Y = b_0 + b_1X + b_2X^2 + \dots + b_kX^k$ (степень полинома k должна быть от 2 до 6);
- степенная — функция регрессии вида $Y = m X^b$;
- экспоненциальная — функция регрессии вида $Y = m e^{bX}$.

На рис. 15.2 показаны точечный график экспериментальной зависимости Y от X и функция регрессии экспоненциального вида. На графике также выводятся уравнение регрессии и значение коэффициента детерминации R^2 . О том, как добавить к диаграмме функцию регрессии (линию тренда) и как задать ее параметры, подробно рассказано в разделе 6.2.1. Недостатком такого способа построения функции регрессии является то, что при необходимости продолжить работу с этой функцией приходится вручную переносить на рабочий лист значения ее коэффициентов.

15.2. Адекватность уравнения регрессии

Статистическая модель. Статистические характеристики уравнения регрессии и коэффициентов функции регрессии обычно определяются при условии, что случайная величина ε из уравнения зависимости $Y(X) = f(X) + \varepsilon$ имеет нормальное распределение с нулевым математическим ожиданием. Другими словами, если наблюдения y_i представимы в виде $y_i = f(x_i) + \varepsilon_i$, то случайные величины ε_i должны быть независимыми и иметь одинаковые нормальные распределения с нулевыми математическими ожиданиями и одинаковыми дисперсиями.

Обычно допускается некоторое отклонение от условия нормальности, но условие равенства дисперсий всех выборочных значений более жестко. Если последнее условие явно не выполняется, то, во-первых, для вычисления коэффициентов

функции регрессии следует применять не стандартный метод наименьших квадратов, а его модификацию — так называемый *взвешенный метод наименьших квадратов* [14], который учитывает неравноточность наблюдений y_i . Во-вторых, основные статистические показатели (см. дисперсионную таблицу из раздела 3.4.3) также вычисляются по измененным формулам [14].

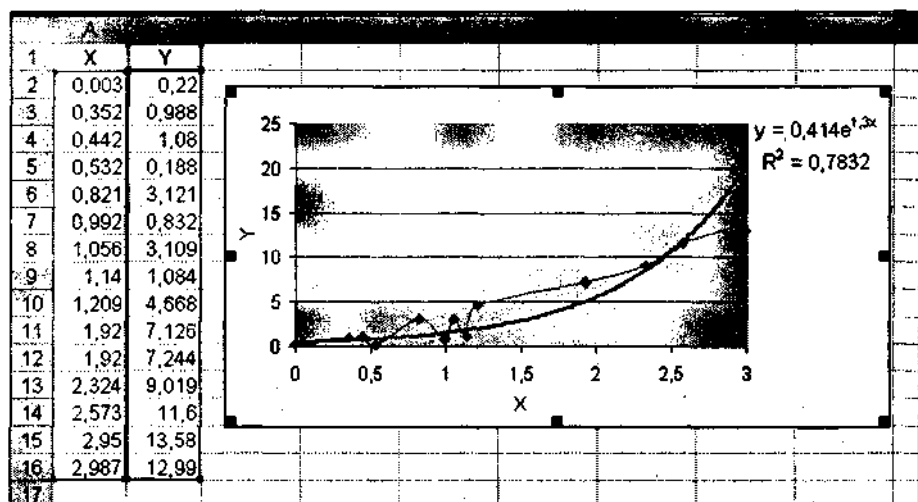


Рис. 15.2. Построение функции регрессии как линии тренда

Далее в этом разделе и в последующих разделах будем предполагать, что выполняется описанная выше статистическая модель.

Критерий проверки адекватности функции регрессии описан в разделе 3.4.3. Напомним, что критерий проверяет нулевую гипотезу $H_0: R^2 = 0$ (R^2 — коэффициент детерминации) против гипотезы $H_1: R^2 \neq 0$. Если нулевая гипотеза отклоняется (с заданным уровнем значимости α), то считается, что функция регрессии статистически значима, т.е. адекватно описывает исходные данные¹. Считаем, что задан уровень значимости α .

Вычисления

1. Вычисляется коэффициент детерминации R^2 . Формула для вычисления приведена в разделе 3.4.3. (Мы не приводим ее здесь, поскольку рассчитываем на применение функции ЛИНЕЙН или средства Регрессия, которые вычисляют этот коэффициент.)

2. Вычисляется критериальная статистика $F = \frac{n-k-1}{k} \cdot \frac{R^2}{1-R^2}$, где n — количество наблюдений y_i , k — количество вычисляемых коэффициентов функции регрессии без свободного члена, т.е. количество столбцов со значениями X в исходных данных. (Другая формула для вычисления статистики F приведена в разделе 3.4.3. Эту статистику также вычисляют функция ЛИНЕЙН и средство Регрессия.)

¹ На практике считается, что если $R^2 > 0,7$, то такое значение значимо априори.

Построение критической области. При условии истинности нулевой гипотезы статистика F имеет F -распределение со степенями свободы k и $(n - k - 1)$.

Определяется критическое значение t как квантиль порядка $1 - \alpha$ F -распределения со степенями свободы k и $(n - k - 1)$. Нулевая гипотеза принимается, если $F \leq t$. В противном случае принимается гипотеза о статистической значимости регрессии.

Практическая реализация в Excel этого критерия не вызывает затруднений, если вычислена критериальная статистика F (с помощью функции ЛИНЕЙН или средства Регрессия). Для реализации критерия необходимо найти только квантиль F -распределения. На рис. 15.3 показан рабочий лист Excel с исходными данными и функцией регрессии, повторяющими рабочий лист на рис. 15.1. Значения статистики F вычислены в ячейках F6 и F15 с помощью функции ЛИНЕЙН, критические значения — в ячейках H8 и H16 с помощью формул, показанных на рис. 15.3. Как и следовало ожидать, регрессия в обоих случаях значима (значения статистик F существенно превышают критические значения).

X	X^2	ln(1+X)	Y	Y = b0 + b1*X + b2*X^2 + b3*ln(1+X)				Уровень значимости	
1,14	1,3	0,833	1,084	b3	b2	b1	b0	0,05	
1,056	1,114	0,749	3,109	5,66757	0,95495	-2,7205	0,10874		
2,987	8,924	2,295	12,99	5,08624	0,38946	4,63022	1,08685		
2,95	8,703	2,272	13,58	R^2	0,96101	1,10996	#N/D	#N/D	
2,573	6,622	2,031	11,6	F	90,3825	11	#N/D	#N/D	
0,532	0,283	0,249	-0,19		334,054	13,552	#N/D	#N/D	
0,442	0,195	0,178	1,08	Критическое значение				3,58743	=FРАСПОБР(K2;3;G6)
0,003	6E-06	6E-06	0,22						
0,352	0,124	0,117	-0,99	Y = b1*X^2 + b2*ln(1+X)					
1,92	3,686	1,545	7,244	b2	b1	b0			
0,821	0,675	0,516	3,121	2,07309	1,00104	0			
1,209	1,462	0,901	4,668		0,84202	0,26353	#N/D		
1,92	3,685	1,544	7,126	R^2	0,9581	1,05852	#N/D		
2,324	5,399	1,856	9,019	F	148,617	13	#N/D		
0,992	0,983	0,685	0,832		333,04	14,566	#N/D		
				Критическое значение				3,80557	=FРАСПОБР(K2;2;G15)

Рис. 15.3. Проверка адекватности регрессии

15.3. Доверительные интервалы и проверка гипотез для коэффициентов функции регрессии

Статистическая модель описана в предыдущем разделе. Предположим, что значения коэффициентов и их среднееквадратические отклонения уже подсчитаны. Среднееквадратические отклонения коэффициентов регрессии вычисляют функция ЛИНЕЙН и средство Регрессия. На рис. 15.4, на котором приведены результаты расчетов с помощью функции ЛИНЕЙН из предыдущего примера, значения среднееквадратических отклонений записаны в ячейках под значениями коэффициентов (диапазоны F4:I4 и F13:G13).

Доверительные интервалы и критерии проверки гипотез о значимости коэффициентов функции регрессии строятся на том основании, что при выполнении условий статистической модели отношение вычисленного коэффициента к его среднеквадратическому отклонению имеет распределение Стьюдента с $(n - k - 1)$ степенью свободы. Для построения доверительных интервалов необходимо вычислить только квантиль t порядка $(1 + \alpha)/2$ этого распределения, где α — заданный уровень значимости. На рис. 15.4 показаны доверительные интервалы для коэффициентов функции регрессии и формулы, по которым вычисляются границы этих интервалов.

K5										=I3-L4*I4	
E	F	G	H	Доверительный уровень							
1	Y = b0 + b1*X + b2*X^2 + b3*ln(1+X)				Доверительный уровень						
2	b3	b2	b1	b0	0,95						
3	5,6676	0,95495	-2,7205	0,10874	Коэффициент k						
4	5,0862	0,38946	4,63022	1,08685	2,5931 =СТЮДРАСПОБР((1-L2)/2,G6)						
5	R^2	0,961	1,10996	#ИД	#ИД	b0	-2,7096	2,927	=I3+L4*I4		
6	F	90,382	11	#ИД	#ИД	b1	-14,727	9,286	=H3+L4*H4		
7		334,05	13,552	#ИД	#ИД	b2	-0,0549	1,9649	=G3+L4*G4		
8						b3	-7,5215	18,857	=F3+F4*L4		
9											
10	Y = b1*X^2 + b2*ln(1+X)				Коэффициент k						
11	b2	b1	b0	2,5326 =СТЮДРАСПОБР((1-L2)/2,G15)							
12	2,0731	1,00104	0	b1	0,3336	1,6685	=G12+L11*G13				
13	0,842	0,26353	#ИД	b2	-0,0594	4,2056	=F12+L11*F13				
14	R^2	0,9581	1,05852	#ИД	=F12-L11*F13						
15	F	148,62	13	#ИД							
16		333,04	14,566	#ИД							
17											

Рис. 15.4. Доверительные интервалы для коэффициентов функции регрессии

Для проверки гипотез о значимости коэффициента функции регрессии вычисляется критериальная статистика как модуль отношения значения этого коэффициента к его среднеквадратическому отклонению. По заданному уровню значимости α вычисляется критическое значение t — квантиль порядка $1 - \alpha$ распределения Стьюдента с $(n - k - 1)$ степенью свободы. Если критериальная статистика меньше критического значения, принимается гипотеза о том, что данный коэффициент равен нулю. В противном случае считается, что коэффициент значимо отличается от нуля. Реализация критерия показана на рис. 15.5, на котором приведены все необходимые формулы. Интересно отметить, что для функции регрессии $Y = b_0 + b_1 X + b_2 X^2 + b_3 \ln(1+X)$ все коэффициенты, кроме одного, оказались незначимо отличными от нуля, несмотря на то что их значения весьма велики по абсолютной величине.

15.4. Доверительный интервал для значения прогноза

Статистическая модель описана в разделе 15.2. Предположим, что подсчитаны значения коэффициентов и остаточная дисперсия s_e^2 (см. раздел 3.4.3). Среднеквадратическое отклонение остатков (корень из остаточной дисперсии) вычис-

ляют функция ЛИНЕЙН и средство Регрессия. На рис. 15.6, на котором приведены результаты расчетов с помощью функции ЛИНЕЙН из предыдущего примера, значения среднеквадратических отклонений остатков записаны в ячейках G5 и G12.

K5		=ABS(I3/I4)									
	F	G	H	I	J	L	M	N	O	P	Q
1	Y = b0 + b1*X + b2*X^2 + b3*ln(1+X)					Уровень значимости					
2	b3	b2	b1	b0		0,05					
3	5,6676	0,95495	-2,7205	0,1087		Критическое значение					
4	5,0862	0,38946	4,63022	1,0868		2,201 =СТЮДРАСПОБР(L2,G6)					
5	0,961	1,10996	#Н/Д	#Н/Д	b0	0,1	не значим	=ЕСЛИ(K5<\$L\$4,"не значим","значим")			
6	90,382	11	#Н/Д	#Н/Д	b1	0,5876	не значим	=ЕСЛИ(K6<\$L\$4,"не значим","значим")			
7	334,05	13,552	#Н/Д	#Н/Д	b2	2,452	значим	=ЕСЛИ(K7<\$L\$4,"не значим","значим")			
8					b3	1,1143	не значим	=ЕСЛИ(K8<\$L\$4,"не значим","значим")			
9						=ABS(F3/F4)					
10	Y = b1*X^2 + b2*ln(1+X)					Коэффициент k					
11	b2	b1	b0			2,1604 =СТЮДРАСПОБР(L2,G15)					
12	2,0731	1,00104		0	b1	3,7986	значим	=ЕСЛИ(K12<\$L\$11,"не значим","значим")			
13	0,842	0,26353	#Н/Д		b2	2,462	значим	=ЕСЛИ(K13<\$L\$11,"не значим","значим")			
14	0,9581	1,05852	#Н/Д			=ABS(F12/F13)					
15	148,62	13	#Н/Д			=ABS(G12/G13)					
16	333,04	14,566	#Н/Д								

Рис. 15.5. Критерий проверки значимости коэффициентов функции регрессии

Чтобы спрогнозировать значение переменной Y в точке x_0 , которая не входит в исходное множество значений $\{x_1, x_2, \dots, x_n\}$ переменной X , используется построенная функция регрессии $f(X)$ и за значение переменной Y в точке x_0 принимается величина $\hat{y} = f(x_0)$. Возможные проблемы, возникающие при прогнозировании, описаны в разделе 3.4.5. Здесь покажем, как построить доверительный интервал для величины $\hat{y} = f(x_0)$ с заданным доверительным уровнем α .

1. Вычисляется значение $\hat{y} = f(x_0)$.

2. Вычисляются среднее \bar{x} значений x_1, x_2, \dots, x_n и сумма $SS_x = \sum_{i=1}^n (x_i - \bar{x})^2$.

3. Вычисляется стандартная ошибка прогноза $s_0 = s_e \sqrt{1 + \frac{1}{n} + \frac{(x_0 - \bar{x})^2}{SS_x}}$.

4. Определяется квантиль t порядка $(1 + \alpha)/2$ распределения Стьюдента с $(n - k - 1)$ степенью свободы.

5. Строится доверительный интервал вида $(\hat{y} - tx_0, \hat{y} + tx_0)$.

На рис. 15.6 показан рабочий лист Excel, на котором построены доверительные интервалы для значений двух функций регрессии, вычисленных при $x_0 = 5$ (ячейка L5). Все формулы, необходимые для построения доверительных интервалов, также показаны на этом рисунке.

	E	F	G	H	I	J	K	L	M	N	O	P	Q
1		$Y = b0 + b1 \cdot X + b2 \cdot X^2 + b3 \cdot \ln(1+X)$					Доверительный уровень						
		b3	b2	b1	b0			0,95					
3		5,6676	0,95495	-2,7205	0,1087		Коэффициент t						
4		5,0862	0,38946	4,63022	1,0868		2,5931	=СТЫОДРАСПОБР((1-L2)/2,G6)					
5		0,961	1,10996	#ИД	#ИД		X0 =	5					
6		90,382	11	#ИД	#ИД		Y(X0) =	20,535	=I3+H3*L5+G3*L5*L5+F3*LN(1+L5)				
7		334,05	13,552	#ИД	#ИД		Стандартное отклонение						
8		$Y = b1 \cdot X^2 + b2 \cdot \ln(1+X)$						1,5874	=G5*КОРЕНЬ(1+1/15+((L5-H15)^2)/H16)				
9		b2	b1	b0			Доверительный интервал						
10		2,0731	1,00104	0			16,419	24,651	=L6+L4*L8				
11		0,842	0,26353	#ИД			=L6-L4*L8						
12		0,9581	1,05852	#ИД			Коэффициент t						
13		148,62	13	#ИД			2,5326	=СТЫОДРАСПОБР((1-L2)/2,G13)					
14		333,04	14,566	#ИД			Y(X0) =	28,741	=G10*L5*L5+F10*LN(1+L5)				
15		Среднее X		1,41468			Стандартное отклонение						
16		Сумма квадратов X		13,1365			1,5138	=G12*КОРЕНЬ(1+1/15+((L5-H15)^2)/H16)					
17				=СРЗНАЧ(A2:A16)			Доверительный интервал						
18		=15*ДИСПР(A2:A16)					24,907	32,574	=L14+L16*L13				
19							=L14-L16*L13						
20													

Рис. 15.6. Доверительные интервалы для значений прогноза

В заключение отметим, что в Excel имеются три функции, которые могут определять значения прогноза без явного вычисления уравнения регрессии.

- ПРЕДСКАЗ вычисляет значения линейной функции регрессии $Y = mX + b$ (см. раздел 4.9.4).
- ТЕНДЕНЦИЯ вычисляет значения полиномиальной функции регрессии, в том числе множественной регрессии (функция описана в разделе 4.9.5). Тип функции регрессии определяется структурой входных данных переменной X так же, как для функции ЛИНЕЙН.
- РОСТ вычисляет значения экспоненциальной функции регрессии, в том числе множественной регрессии (функция описана в разделе 4.9.7). Тип функции регрессии определяется структурой входных данных переменной X так же, как для функции ЛГРФПРИБЛ.

Все три функции могут использоваться в формулах массивов и, таким образом, могут вычислять не только отдельные значения, но и массивы значений. Их удобно применять для вычисления остатков, т.е. разностей между исходными значениями переменной Y и значениями, вычисленными по функции регрессии. На рис. 15.7 показан рабочий лист Excel и соответствующие формулы, с помощью которых вычисляются прогнозируемые значения переменной Y и остатки для исходных данных и двух функций регрессии, рассмотренных в предыдущих примерах.

H2		{=Y-ТЕНДЕНЦИЯ(Y:B2:C16;B2:C16;0)}								
X	X^2	ln(1+X)	Y	Y1	Остатки 1	Y2	Остатки 2			
1,14	1,3	0,8329	1,084	2,989	-1,984868	3,028	-1,944075			
1,056	1,114	0,7488	3,109	2,545	0,564387	2,668	0,441284			
2,997	8,924	2,2948	12,89	13,51	-0,518296	13,691	-0,698781			
2,95	8,703	2,2724	13,58	13,27	0,307587	13,423	0,157582			
2,573	6,622	2,031	11,6	10,94	0,653212	10,839	0,756417			
0,532	0,283	0,2482	-0,189	0,344	-0,532289	0,7998	-0,988108			
0,442	0,195	0,1785	1,08	0,104	0,975302	0,5858	0,513988			
0,003	6E-06	6E-06	0,22	0,102	0,117878	2E-05	0,219921	{=Y-ТЕНДЕНЦИЯ(Y:B2:C16;B2:C16;0)}		
0,352	0,124	0,1187	-0,988	-0,089	-0,91883	0,3657	-1,353448			
1,92	3,686	1,5446	7,244	7,18	0,084578	6,8919	0,35222			
0,821	0,675	0,5156	3,121	1,44	1,680942	1,7441	1,377304			
1,209	1,462	0,901	4,668	3,322	1,345966	3,3316	1,336393			
1,92	3,685	1,5444	7,126	7,159	-0,032752	6,891	0,234882			
2,324	5,399	1,8582	9,019	8,484	-0,44459	8,2532	-0,234253			
0,982	0,883	0,8847	0,832	2,231	-1,398245	2,4036	-1,57126			
{=ТЕНДЕНЦИЯ(Y:A2:C16;A2:C16)}							{=ТЕНДЕНЦИЯ(Y:B2:C16;B2:C16;0)}			
{=Y-ТЕНДЕНЦИЯ(Y:A2:C16;A2:C16)}										

Рис. 15.7. Вычисление прогнозируемых значений и остатков

Литература

1. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: исследование зависимостей. — М. : Финансы и статистика, 1985.
2. Айвазян С.А., Мхитарян В.С. Прикладная статистика и основы эконометрики. — М. : ЮНИТИ-ДАНА, 2001.
3. Антон Г. Анализ таблиц сопряженности. — М. : Финансы и статистика, 1982.
4. Бальшее Л.Н., Смирнов Н.В. Таблицы математической статистики. — 3-е изд. — М. : Наука, 1983.
5. Боровков А.А. Математическая статистика. — М. : Наука, 1984.
6. Гихман И.И., Скороход А.В., Ядренко М.И. Теория вероятностей и математическая статистика. — 2-е изд. — К. : Вища шк., 1988.
7. Ивченко Г.И., Медведев Ю.И. Математическая статистика. — М. : Высш. шк., 1984.
8. Королук В.С., Портенко Н.И., Скороход А.В., Турбин А.Ф. Справочник по теории вероятностей и математической статистике. — К. : Наук, думка, 1978.
9. Ликеш И., Ляга Й. Основные таблицы математической статистики. — М. : Финансы и статистика, 1985.
10. Макарова Н.В., Трофимец В.Я. Статистика в Excel. — М. : Финансы и статистика, 2002.
11. Минько А.А., Петунин Ю.И. Сходимость метода наименьших квадратов в равномерной метрике // Сиб. матем. ж. — 1990. — № 2.
12. Мур Дж., Уэдерфорд Л. Экономическое моделирование в Microsoft Excel. — 6-е изд. — М. : Издат. дом "Вильямс", 2004.
13. Петрович М.Л., Давидович М.И. Статистическое оценивание и проверка гипотез на ЭВМ. — М. : Финансы и статистика, 1989.
14. Поллард Дж. Справочник по вычислительным методам статистики. — М. : Финансы и статистика, 1982.
15. Сигел Э.Ф. Практическая бизнес-статистика. — М. : Издат. дом "Вильямс", 2002.
16. Соболев И.М. Численные методы Монте-Карло. — М. : Наука, 1973.
17. Справочник по прикладной статистике. Т.1. / Под ред. Э. Лойда, У. Ледермана. — М. : Финансы и статистика, 1989.
18. Справочник по прикладной статистике. Т.2. / Под ред. Э. Лойда, У. Ледермана. — М. : Финансы и статистика, 1990.
19. Тюрин Ю.Н., Макаров А.А. Статистический анализ данных на компьютере. — М. : Инфра-М, 1998.
20. Уокенбах Дж., Андердал Б. Excel 2002. Библия пользователя. — М. : Диалектика, 2002.

21. Ханк Дж.Э., Райте А.Дж., Уичерни Д.У. Бизнес-прогнозирование^ — М. : Издат. дом "Вильяме", 2003.
22. Холлендер М., Вулф Д. Непараметрические методы статистики. — М. : Финансы и статистика, 1983.
23. Хьюберт П. Робастность в статистике. — М. : Мир, 1984.
24. Шеффе Г. Дисперсионный анализ. — М. : Наука, 1980.

Предметный указатель

р

F-распределение, 41

J^{\wedge}

Адекватность уравнения регрессии, 420

Алгебра случайных событий, 22

Анализ

дисперсионный, 94; 408

корреляционный, 81 ;»382

регрессионный, 88; 181; 417

Фурье, 189

Анализ зависимостей между

классификационными переменными, 86

количественными переменными, 81

порядковыми переменными, 83

В

Вариационной ряд, 59; 65

Вероятность

ошибки второго рода, 66

ошибки первого рода, 66

условная, 22

Винзоризация, 258

Выборка, 49

размах, 65

широта, 65

Выборочная дисперсия, 50

Выборочное

пространство, 49

среднее, 50

Выборочный момент, 50

Выбросы, 250; 288

Г

Генеральная совокупность, 49

Гипотеза

альтернативная, 65

конкурирующая, 65

нулевая, 65

статистическая, 65

Гистограмма, 141; 286

с перекрытием, 215

Д

Дециль, 27

Диаграмма, 206

планки погрешностей, 210

Диалоговое окно

Анализ данных, 146; 147

Анализ Фурье, 191

Выборка, 160

Генерация случайных чисел, 155

Гистограмма, 152

Двухвыборочный F-тест для дисперсий, 172

Двухвыборочный t-тест с одинаковыми дисперсиями, 165

Двухвыборочный t-тест с различными дисперсиями, 167

Двухвыборочный z-тест для средних, 163

Двухфакторный дисперсионный анализ без повторений, 178

Двухфакторный дисперсионный анализ с повторениями, 175

Добавление ограничения, 221

Ковариация, 181

Корреляция, 179

Линия тренда, 207

Однофакторный дисперсионный анализ, 173

Описательная статистика, 147; 149; 151

Параметры поиска решения, 221

Парный двухвыборочный t-тест для средних, 170

Поиск решения, 221; 227

Присвоение имени, 196; 199; 200

Ранг и персентиль, 161

Регрессия, 182

- Результаты поиска решения, 223
- Скользящее среднее, 188
- Создать имена, 200
- Специальная вставка, 232
- Таблица подстановки, 232
- Формат линии тренда, 210
- Формат планок погрешностей, 211
- Формат ряда данных, 210; 212; 217; 270
- Формат ячеек, 254
- Экспоненциальное сглаживание, 189
- Дисперсионное отношение Фишера, 75; 376
- Дисперсионный анализ, 94
 - двухфакторный, 97; 175; 177; 409
 - многофакторный, 95
 - модель с постоянными факторами, 95
 - модель смешанная, 95
 - модель со случайными факторами, 95
 - однофакторный, 95; 173
 - статистическая модель, 94
 - таблица, 96
 - факторы, 94
 - Фридмана, 411
- Дисперсия, 26
 - выборочная, 112
 - интервальные оценки, 57
 - точечная оценка, 56
- Доверительная область, 53
- Доверительные границы, 52
- Доверительные интервалы
 - для дисперсий, 310; 315
 - для квантилей, 333
 - для математического ожидания, 307; **312**
- Доверительный интервал, 52
 - для значения прогноза, 423
 - для коэффициента корреляции, **393**
 - для коэффициента корреляции нормальной совокупности, 394
 - для коэффициентов функции регрессии, 422
 - для математических ожиданий нескольких совокупностей, 401
 - для математического ожидания, 140
 - для отношения дисперсий нормальных совокупностей, 366
 - для разности двух биномиальных вероятностей, 367
 - для разности математических ожиданий, 400
 - для разности средних нормальных совокупностей, 364; 365

3

- Задача
 - восстановления значений, 93
 - экстраполяции, 94
- Задачи статистического анализа зависимостей, 79
- Закон больших чисел, 21
- Закон распределения, 23

.И

- Имена
 - диапазона, 197
 - переопределение, 200
 - создание, 199
 - ячеек, 197
- Индекс корреляции, 83; 92; 183
- Интервал
 - модальный, 285

К

- Квантиль, 27; 65
 - доверительные интервалы, 333
- Квартиль, 27; 104
- Ковариация, 28
 - выборочная, 136
- Корреляционный анализ, 81
- Коэффициент
 - асимметрии, 26; 58; 110; 291
 - асимметрии Пирсона, 27; 28
 - детерминации, 83; 92; 96; 130; 184
 - конкордации, 85; 389
 - корреляции, 28; 81; 137; 139; 170; 384
 - корреляции выборочный, 137
 - корреляции Кендалла, 85; 386
 - корреляции множественный, **183**
 - корреляции Пирсона, 138
 - корреляции ранговый, 84
 - корреляции Спирмена, 84; 385

- согласованности, 85; 389
- эксцесса, 26; 58; 111; 291
- Критерии**
 - независимости, 382
 - однородности, 349
- Критерии проверки статистических гипотез**
 - непараметрические, 67
 - робастные, 67
 - свободные от распределений, 67
 - устойчивые, 67
- Критерий**
 - χ^2 (Пирсона), 76
 - Ансари-Бредли проверки гипотезы о равенстве дисперсий, 378
 - Бартлетта проверки равенства нескольких дисперсий, **376**
 - Бартлетта, 95
 - Беренса-Фишера, 73
 - Беренса-Фишера проверки гипотезы о равенстве математических ожиданий, 370
 - знаков, 344; 405
 - знаковых рангов Уилкоксона, 346
 - Колмогорова, 77; 304
 - Краскала-Уоллиса, 357
 - медианы, 350
 - минимума суммы квадратов, 88
 - множественных сравнений Шеффе, 96; 373; 415
 - отклонения от распределения Пуассона, 296
 - отклонения распределения от нормального, 293
 - Пирсона, 297
 - проверки статистической гипотезы, **66**
 - серий Вальда-Вольфовица, **359**
 - Смирнова, 362
 - согласия хи-квадрат, 297
 - Стьюдента модифицированный, **96**
 - Стьюдента парный, **404**
 - Уилкоксона, 407
 - Уилкоксона-Манна-Уитни, 74; 202; 355
 - Фишера, 95; 172
 - Фишера проверки **равенства** дисперсий, 375
 - Фишера-Беренса, 167
 - Фридмана, 411
 - хи-квадрат, 127; 226; 297; 299; 360
- Критерий независимости**
 - для двумерных нормальных совокупностей, 384
 - для многомерных выборок, 389
 - на основе коэффициента корреляции Кендалла, 386
 - на основе коэффициента корреляции Спирмена, 385
 - на основе преобразования Фишера, 383
 - на основе таблиц сопряженности, **390**
 - хи-квадрат, 390
- Критерий проверки**
 - значения дисперсии нормальной совокупности, 337
 - значения коэффициента корреляции, **396**
 - равенства двух коэффициентов корреляции, 397
 - равенства нескольких коэффициентов корреляции, 399
- Критерий проверки гипотез**
 - о значении коэффициента корреляции, 396
 - о значении математического ожидания, 124
 - о значении медианы, 343
 - о значении параметра биномиального распределения, 341
 - о значении параметра показательного распределения, 339
 - о параметрах нормального распределения, 335
 - о равенстве биномиальных вероятностей, 380
 - о равенстве математических ожиданий, 368
 - о равенстве нескольких математических ожиданий, 371
- Критерий проверки гипотезы**
 - о равенстве математических ожиданий, 124
 - о равенстве математических ожиданий для нормальных совокупностей, 71
- Критерий проверки значения**
 - дисперсии нормальной совокупности, 70
 - математического ожидания нормальной совокупности, 69; 335

Критерий проверки статистической гипотезы, 66
критическая область, 66
критические значения, 66

Критерий Стьюдента
парный, 194
проверки гипотезы о равенстве математических ожиданий, 369
проверки гипотезы о равенстве математических ожиданий для зависимых нормальных совокупностей, 73
проверки гипотезы о равенстве математических ожиданий для нормальных совокупностей, 71; 72

Критерий Фишера
проверки равенства дисперсий, 75; 126

Критическая точка
левосторонняя, 67
правосторонняя, 67

Л

Линия тренда, 207; 420
параметры, 209
форматирование, 209

М

Массив, 193
констант, 193; 196
Математическое ожидание
интервальные оценки, 55
точечная оценка, 54
Матрица
ковариационная, 180
корреляционная, 179; 246
Медиана, 27; 106
вычисление, 284
точечная оценка, 59

Метод
быстрого преобразования Фурье, 190
наименьших квадратов, 89; 417
Неймана, 242
непараметрические, 65
обратных функций, 234
отбора, 242

свободные от распределения, 65
скользящего среднего, 187
суперпозиций, 238
цензурирования Тьюки, 257

Мода, 27; 141

вычисление, 285

Моделирование случайных величин, 229

зависимых, 245
метод Неймана, 242
метод обратных функций, 234
метод отбора, 242
метод суперпозиций, 238
многомерных, 244

Модель статистических зависимостей, 78

Мощность критерия, 66

Н

Надстройка

Пакет анализа, 146

Поиск решения, 217

Начальные моменты

точечные оценки, 58

Неравенство

Гаусса, 28; 55; 251; 308

Маркова, 27

Пика, 28

Чебышева, 27; 55; 251; 252; 308

Несмещенность оценки, 50

О

Область

двухсторонняя критическая, 67

критическая, 66

левосторонняя критическая, 67

непринятия гипотезы, 66

правосторонняя критическая, 67

Однофакторный дисперсионный анализ, 371

Отклонение нормированное среднее

абсолютное, 292

Оценка

асимптотически несмещенная, 51

интервальная, 50; 52

квантилей, 65

несмещенная, 50

параметра распределения Бернулли, 61

параметра распределения Пуассона, 63
параметров нормального
 распределения, 59
состоятельная, 51
точечная, 49
точечные, вычисление, 278; 283
эффективная, 51
Оценки параметров
 гамма-распределения, 319
 геометрического распределения, 331
 логарифмически нормального
 распределения, 317
 нормального распределения, 312
 показательного распределения, 318
 равномерного распределения, 323
 распределения Бернулли, 324
 распределения Пуассона, 329
Ошибка
 второго рода, 66
 первого рода, 66

П

Переменные
 классификационные, 79
 количественные, 79
 номинальные, 79
 ординальные, 79
 порядковые, 79
Пирсона коэффициент корреляции, 138
Планки погрешностей, 210
Плотность вероятности, 25
Порядковые статистики, 65
Построение
 гистограмм, 212; 267; 273
 полигонов, 267; 273
 пробит-графика, 288
 функции регрессии, 90
 функций распределения, 212
 эмпирических функций
 распределения, 267
Правило трех сигм, 39
Преобразование
 арксинуса, 63; 328
 квадратного корня, 263
 логарифмическое, 265
 стандартизирующее, 267

Фишера, 82; 139; 383; 393
Фурье дискретное, 189
Энскомба, 63
Пробит-график, 288
 построение, 288
Проверка гипотез
 для коэффициентов функции
 регрессии, 422
Прогнозирование, 93
Процентиль, 27; 105

Р

Размах, 279
 выборки, 65
 интерквартильный, 279
Разности кумулятивные, 304
Ранг, 65; 107
 вычисление, 202
 процентный, 106
Распределение
 безгранично делимое, 245
 Бернулли, 32; 154; 230; 324
 бета, 43; 114; 121; 234
 биномиальное, 33; 115; 123; 154; 230; 341
 Вейбулла-Гнеденко, 44; 115
 гамма, 44; 116; 121; 234; 319
 геометрическое, 34; 331
 гипергеометрическое, 35; 116
 двумерное, 28
 дискретное, 155; 230
 дискретное равномерное, 144
 Кендалла, 387
 Колмогорова-Смирнова, 77; 304
 Краскала-Уоллиса, 358
 логарифмически нормальное, 42; 117;
 121; 234; 252; 317
 логнормальное, 42
 Манна-Уитти, 75; 357
 модельное, 154
 нормальное, 38; 117; 122; 154; 230; 234
 одномодальное, 27; 55; 252; 308
 отрицательное биномиальное, 35; 117
 Паскаля, 35; 117
 показательное, 37; 119; 318; 339
 Пуассона, 34; 118; 154; 230; 263; 296;
 329; 354
 равномерное, 36; 144; 154; 230; 323

равномерное дискретное, 32
 случайных величин, 23
 Смирнова, 362
 Снедекора, 41; **114; 120; 234; 342**
 Спирмена, 385
 стандартное нормальное, 38; 117; 122
 Стьюдента, 40; 118; 122; 169; 234; 295;
 309; 317; 336; 347; 370; 384
 треугольное, 37
 Уилкоксона, 407
 Фридмана, 413
 хи-квадрат, 39; 119; **122; 234; 264; 297;**
 330; 338; 351; 361
 экспоненциальное, 37; 119
 Распределения Пирсона, 45; **291**
 Регрессионный анализ, 88
 Регрессия, 182; 417
 линейная, 91
 множественная, 90
 нелинейная, 89
 построение функции, 90
 проверка адекватности, 91
 статистические характеристики, 92
 уравнение, 88
 функция, 88
 функция полиномиальная, 89
 экспоненциальная, 134; 135

С

Система
 нормальных уравнений, **418**
 Случайная величина
 асимптотически нормальная, 32
 дециль, 27
 дискретная, 23
 дисперсия, 26
 квантиль, 27
 квартиль, 27
 коэффициент асимметрии, 26
 коэффициент эксцесса, 26
 линейное преобразование, 30
 математическое ожидание, 26
 медиана, 27
 мода, 27
 моменты, 26
 непрерывная, 25

нормирование, 30
 процентиль, 27
 стандартизованная, 30
 центральные моменты, 26
 числовые характеристики, 25
 Случайное событие, 20
 Случайный опыт, 20
 Состоятельность оценки, 51
 Среднее
 арифметическое, 109
 гармоническое, 109
 геометрическое, 109
 скользящее, 187
 Средство
 Анализ Фурье, 189
 Выборка, 160
 Генерация случайных чисел, 154; 225;
 229; 230; 235; 301; 309; 336; 354
 Гистограмма, 151; 226; 270; 278
 Двухвыборочный F-тест для дисперсий,
 127; 172; 376
 Двухвыборочный t-тест с одинаковыми
 дисперсиями, 126; 165
 Двухвыборочный t-тест с различными
 дисперсиями, 73; 126; 167; 371
 Двухвыборочный z-тест для средних,
 161; 369
 Двухвыборочный z-тест с одинаковыми
 дисперсиями, 370
 Двухфакторный дисперсионный анализ
 без повторений, 177; 411
 Двухфакторный дисперсионный анализ
 с повторениями, 175
 Ковариация, 180
 Корреляция, 179
 Однофакторный дисперсионный анализ,
 173; 373
 Описательная статистика, **149; 279**
 Парный двухвыборочный t-тест для
 средних, 125; 169
 Подбор параметра, 236
 Поиск решения, 236; 237; **418**
 Ранг и перцентиль, 161; 202
 Регрессия, 93; 181; 418; 422
 Скользящее среднее, 187
 Экспоненциальное сглаживание, 188

Средство Поиск решения
 подбор параметров, 224
 подбор параметров распределения, 225
 поиск безусловного оптимума, 224
 поиск допустимого решения, 224
 поиск оптимума, 224
 решение системы линейных
 алгебраических уравнений, 225
 Статистика, 50
 критериальная, 68
 оценивания дисперсии, 56
 оценивания коэффициента
 асимметрии, 58
 оценивания коэффициента эксцесса, 58
 оценивания математического
 ожидания, 54
 оценивания медианы, 59
 оценивания моментов, 58
 Статистики
 порядковые, 65
 ранговые, 65
 Статистическая модель, 54
 Сумма квадратов
 остатков, 130
 регрессии, 130
 Суммы случайных величин, 30

Т

Таблица
 дисперсионная, 91; 96
 дисперсионного анализа, 96
 сопряженности, 391
 частотная, 268
 Таблица
 подстановки, 232
 сопряженности, 86
 Теорема
 сложения вероятностей, 22
 умножения вероятностей, 22
 центральная предельная, 31

У

Уравнение регрессии, 88
 Уровень
 доверительный, 52

значимости, 52
 значимости критерия, 66
 Условное форматирование, 253

Ф

Фишера дисперсионное отношение, 75
 Формула
 массива, 141; 193; 196
 Стерджесса, 152; 273
 Функции
 FPАСп, 114
 FPАСпОВР, 120; 234
 TANH, 394
 ZТЕСТ, 124; 164; 336
 БЕТАОБР, 121; 234
 БЕТАРАСП, 114
 БИЗВЛЕЧЬ, 352
 БИНОМРАСП, 115
 ВЕЙБУЛЛ, 115
 вероятностей, 24
 ВЕРОЯТНОСТЬ, 140
 ВПР, 236; 352; 364
 вычисления выборочной дисперсии
 и отклонения, 111
 вычисления геометрических
 характеристик распределения, 110
 вычисления значений функций
 распределения, 113
 вычисления ковариации, 136
 вычисления коэффициента
 корреляции, 136
 вычисления средних, 109
 гамма Эйлера, 142
 ГАММАНЛОГ, 142
 ГАММАОБР, 121; 234; 319; 340
 ГАММАРАСП, 116
 ГИПЕРГЕОМЕТ, 116
 ДВССЫЛ, 302
 ДИСП, 112
 ДИСПА, 112
 ДИСПР, 112; 338
 ДИСПРА, 112
 ДОВЕРИТ, 60; 140; 313
 ЕНД, 352
 ЗНАК, 388
 КВАДРОТКЛ, 112

КВАРТИЛЬ, 104
 КВПИРСОН, 138
 КОВАР, 136; 180
 КОРРЕЛ, 137; 180; 384
 КРИТБИНОМ, 123
 ЛГРФПРИБЛ, 134; 420; 425
 ЛИНЕЙН, 93; 129; 132; 138; 418; 422
 ЛОГНОРМОБР, 121; 234
 ЛОГНОРМРАСП, 117
 МАКС, 103; 392
 МАКСА, 103
 МЕДИАНА, 106
 МИН, 103
 МИНА, 103
 МОБР, 204
 МОДА, 141; 285
 МОПРЕД, 204
 МУМНОЖ, 204
 НАИБОЛЬШИЙ, 103
 НАИМЕНЬШИЙ, 103
 НАКЛОН, 131; 420
 нелинейной регрессии, 89
 НОРМАЛИЗАЦИЯ, 142
 НОРМОБР, 122; 234
 НОРМРАСП, 117
 НОРМСТОБР, 122; 234; 288
 НОРМСТРАСП, 117
 обратные к функциям распределения, 119
 определения экстремальных значений
 выборки, 102
 ОСТАТ, 202
 от случайных величин, 29
 ОТРБИНОМРАСП, 117
 ОТРЕЗОК, 131; 420
 ПЕРЕСТ, 143
 ПЕРСЕНТИЛЬ, 105
 ПИРСОН, 137
 построения уравнения регрессии, 128
 ПРЕДСКАЗ, 133; 134; 425
 проверки статистических критериев, 123
 ПРОЦЕНТРАНГ, 106
 ПУАССОН, 118; 298
 работы с порядковыми статистиками, 104
 РАНГ, 107; 202; 333; 347; 352; 379;
 386; 390
 распределения частная, 28
 распределения, 24
 регрессии, 88; 129
 регрессии полиномиальная, 89
 РОСТ, 135; 425
 СКОС, 110
 СЛУЧМЕЖДУ, 144; 230
 СЛЧИС, 144; 229
 СРГАМ, 109
 СРГЕОМ, 109
 СРЗНАЧ, 109
 СРЗНАЧА, 109
 СРОТКЛ, 113
 СТАНДОТКЛОН, 112
 СТАНДОТКЛОНА, 112
 СТАНДОТКЛОНП, 113
 СТАНДОТКЛОНПА, 113
 СТОИУУХ, 132; 290
 СТРОКА, 302
 СТЬЮДРАСП, 118
 СТЬЮДРАСПОБР, 122; 234; 235; 314; 366
 СУММЕСЛИ, 200
 СУММКВ, 204
 СУММКВРАЗН, 205; 386
 СУММПРОИЗВ, 205
 СУММРАЗНКВ, 206
 СУММСУММКВ, 206
 СЧЁТ, 143; 392
 СЧЁТЗ, 143
 ТЕНДЕНЦИЯ, 133; 135; 425
 ТТЕСТ, 124; 167; 169; 171
 УРЕЗСРЕДНЕЕ, 110
 ФИШЕР, 139; 384; 394
 ФИШЕРОБР, 139; 394
 ФТЕСТ, 126; 173
 ХИ2ОБР, 122; 234; 300; 315; 330
 ХИ2РАСП, 119
 ХИ2ТЕСТ, 127; 298; 300; 303
 ЧАСТОТА, 141; 277; 298
 ЭКСПРАСП, 119; 319
 ЭКСЦЕСС, 111
 Функция регрессии
 построение, 418

Ц

Цензурирование, 250
 метод Тьюки, 257

на основе доверительных
интервалов, 251
непараметрическое, 257
Центральная предельная теорема, 31
Центральные моменты
точечные оценки, 58

Ч

Частоты, 268
накопленные, 268; 288
накопленные относительные, 268
Частота события, 21

Частоты, 268
накопленные, 268
относительные, 268
относительные накопленные, 268

Ш

Широта выборки, 65

Э

Экспоненциальное сглаживание, 188
Эффективность оценки, 51

Научно-популярное издание

Александр Александрович Минько

Статистический анализ в MS Excel

Литературный редактор *Л.Н. Красножон*

Верстка *В.И. Бордюк*

Художественный редактор *В.Г. Павлютин*

Корректоры *З.В. Александрова, Л.А. Гордиенко,
О.В. Мишутина*

Издательский дом "Вильямс".
101509, Москва, ул. Лесная, д. 43, стр. 1.

Подписано в печать 20.09.2004. Формат 70X100/16.
Гарнитура Times. Печать офсетная.
Усл. печ. л. 36,12. Уч.-изд. л. 27,5.
Тираж 3000 экз. Заказ № 696.

Отпечатано с диапозитивов в ФГУП "Печатный двор"
Министерства РФ по делам печати,
телерадиовещания и средств массовых коммуникаций.
197110, Санкт-Петербург, Чкаловский пр., 15.