6 ОЦЕНИВАНИЕ ВЗАИМОСВЯЗИ ПАРАМЕТРОВ С ПРИМЕНЕНИЕМ ДИСПЕРСИОННОГО АНАЛИЗА

Основоположник дисперсионного анализа Р. А. Фишер в 1938 г. определил его сущность как «отделение дисперсии, приписываемой одной группе причин, от дисперсии, приписываемой другим группам». В современном понимании задача дисперсионного анализа состоит в исследовании влияния факторов на изменчивость отклика.

В ходе такого исследования выборочная дисперсия отклика раскладывается на составляющие – случайную (обусловленную действием возмущений, ошибками наблюдений и т. п. причинами) и систематические (обусловленные действием того или иного фактора). Каждая из систематических (объясненных) составляющих представляет собой оценку дисперсии генеральной совокупности соответствующего фактора. Чтобы решить, значимо ли влияние конкретного фактора на изменчивость отклика, необходимо сравнить его выборочную дисперсию со случайной составляющей общей дисперсии отклика.

При разработке аппарата дисперсионного анализа были сделаны следующие допущения:

1. Случайные ошибки наблюдений имеют нормальное распределение.
2. Факторы влияют на изменения только среднего выборочного значения отклика, а дисперсия наблюдений остается постоянной.
3. Все эксперименты равноточные.

6.1.Однофакторный дисперсионный анализ

Пусть необходимо оценить влияние на отклик единичного фактора X, который принимает k различных значений (уровней фактора). На i-м уровне производится n_i наблюдений, а общее число опытов $N = n_1 + n_2 + ... + n_k$.

Предполагается, что результат любого наблюдения можно представить в виде модели:

$$y_{ij} = \mu + d_i + \varepsilon_{ij},$$

где μ - суммарный эффект во всех опытах;

d_i - эффект фактора на i-м уровне;

ε_{ij} - ошибка измерения на i-м уровне.

Предполагается также, на фиксированном уровне фактора результаты наблюдений нормально распределены относительного среднего значения $\mu + d_i$ с общей дисперсией σ^2.

Необходимо проверить нулевую гипотезу о равенстве средних значений отклика на различных уровнях фактора:

$$H_0 : m_1 = m_2 = ... = m_k = m.$$

6.1.1.Однофакторный дисперсионный анализ с равным числом повторений

Наиболее простые расчеты получаются при равном числе наблюдений на каждом уровне, т. е. в случае $n_1 = n_2 = ... = n_k = n$. В таком случае исходные данные для дисперсионного анализа представляются в виде таблицы (табл. 6.1), в которой также подсчитываются итоги по столбцам A_i.

Затем вычисляется:

сумма квадратов всех наблюдений:

$$SS_1 = \sum_{i=1}^{k} \sum_{j=1}^{n} y^2{}_{ij};$$

средняя сумма квадратов итогов:

$$SS_2 = \frac{1}{n} \sum_{i=1}^{k} A_i^2;$$

средний квадрат общего итога:

Таблица 6.1

Исходные данные однофакторного дисперсионного анализа с равным числом повторений

Номер наблюдения	Уровни фактора X			
	x_1	x_2	…	x_k
1	y_{11}	y_{21}	…	y_{k1}
2	y_{12}	y_{22}	…	y_{k2}
…	…	…	…	…
n	y_{1n}	y_{2n}	…	y_{kn}
Итоги	$A_1 = \sum_{j=1}^{n} y_{1j}$	$A_2 = \sum_{j=1}^{n} y_{2j}$	…	$A_k = \sum_{j=1}^{n} y_{kj}$

$$SS_3 = \frac{1}{N}\left(\sum_{i=1}^{k} A_i\right)^2;$$

сумма квадратов для столбца:

$$SS_A = SS_2 - SS_3;$$

общая сумма квадратов:

$$SS_{общ} = SS_1 - SS_3;$$

остаточная сумма квадратов:

$$SS_{ост} = SS_1 - SS_2.$$

Дисперсия отклика, обусловленная фактором (объясненная дисперсия):

$$s_X^2 = \frac{SS_A}{k-1}.$$

Дисперсия отклика, обусловленная случайными причинами:

$$s_{ош}^2 = \frac{SS_{ост}}{k(n-1)}.$$

Результаты расчетов представляются в виде специальной таблицы (табл. 6.2).

Таблица 6.2.

Однофакторный дисперсионный анализ
с равным числом повторений

Источник дисперсии	Число степеней свободы	Сумма квадратов	Средний квадрат	Математическое ожидание среднего квадрата
X	$k-1$	SS_A	s_X^2	$n\sigma_X^2+\sigma_{ош}^2$
Остаток	$k(n-1)$	$SS_{ост}$	$s_{ош}^2$	$\sigma_{ош}^2$
Общая сумма	$kn-1$	$SS_{общ}$	$\dfrac{SS_{общ}}{kn-1}$	

Сформулированная выше нулевая гипотеза о равенстве средних значений отклика на всех уровнях фактора справедлива, если отличие между s_X^2 и $s_{ош}^2$ является незначимым. Действительно, если $s_X^2 = s_{ош}^2$, то это означает, что вся вариация отклика обусловлена случайными причинами, а не воздействием фактора X. Если же влияние фактора существенно, то часть дисперсии отклика, обусловленная вариацией фактора (s_X^2) должна быть больше, чем часть дисперсии отклика, обусловленная случайными причинами ($s_{ош}^2$).

В математической статистике для сравнения дисперсий используют распределение и критерий Фишера. Распределение Фишера представляет собой распределение случайной величины

$$F = \left(\frac{s_1^2}{\sigma_1^2}\right) : \left(\frac{s_2^2}{\sigma_2^2}\right),$$

которое зависит только от числа степеней свободы ν_1 и ν_2. Нулевой гипотезой является предположение о равенстве сравниваемых дисперсий ($H_0 : \sigma_1^2 = \sigma_2^2$). В таком случае $\sigma_1^2 / \sigma_2^2 = 1$ и распределение Фишера можно непосредственно использовать для оценивания отношения выборочных оценок дисперсии s_1^2 / s_2^2. Это

отношение принято называть критерием Фишера. Для его расчета в числителе следует использовать большее значение (т. е. должно быть $s_1^2 > s_2^2$). Различие между дисперсиями следует считать значимым, если

$$F = \frac{s_1^2}{s_2^2} > F[\alpha;\nu_1;\nu_2].$$

Применительно к рассматриваемому случаю дисперсионного анализа влияние фактора X на изменчивость отклика следует признавать значимым, если выполняется условие:

$$F = \frac{s_X^2}{s_{ош}^2} > F[\alpha;\nu_1;\nu_2].$$

Если данное условие выполняется, т. е. влияние фактора на отклик является значимым, степень такого влияния может быть рассчитана по формуле:

$$\sigma_X^2 \approx \frac{s_X^2 - s_{ош}^2}{n}.$$

6.1.2.Однофакторный дисперсионный анализ с различным числом повторений

Данный случай имеет место, если на одном или нескольких уровнях факторов число наблюдений отличается от числа наблюдений на других уровнях. Пусть на уровне x_i проведено n_i наблюдений. Общее число наблюдений будет равно:

$$N = \sum_{i=1}^{k} n_i .$$

Итоги по столбцам рассчитываются аналогично, только в каждом столбце суммирование необходимо производить по числу наблюдений, соответствующих данному уровню фактора:

$$A_i = \sum_{j=1}^{n_i} y_{ji} .$$

Сумма квадратов всех наблюдений будет равна:

$$SS_1 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij}^2 .$$

Сумма средних квадратов итогов по столбцам должна рассчитываться следующим образом:

$$SS_2 = \sum_{i=1}^{k} \frac{A_i^2}{n_i} .$$

Остальные величины, необходимые для выполнения дисперсионного анализа, рассчитываются также, как и при одинаковом числе повторений на каждом уровне. Остается без изменений и условие значимости влияния фактора на отклик. Однако степень такого влияния, если оно признано значимым, вычисляется по иной формуле:

$$\sigma_X^2 = \frac{N(k-1)}{N^2 - N}\left(s_X^2 - s_{ош}\right).$$

6.2.Двухфакторный дисперсионный анализ

Задача двухфакторного дисперсионного анализа заключается в оценивании влияния на отклик сразу двух факторов X_1 и X_2, каждый из которых может иметь различное число уровней варьирования (например, k для X_1 и m для X_2). Если при каждом сочетании уровней обоих факторов проводится n наблюдений, то общее число наблюдений будет $N = nkm$. В общем случае результат каждого наблюдения можно представить в виде:

$$y_{ijq} = \mu + \alpha_i + \beta_j + \alpha_i\beta_j + \varepsilon_{ijq},$$

где μ - общее среднее;

α_i - эффект фактора X_1 на i-м уровне ($i = 1,2,...,k$);

β_j - эффект фактора X_2 на j-м уровне ($j = 1,2,...,m$);

$\alpha_i\beta_j$ - эффект взаимодействия факторов;

ε_{ijq} - ошибка воспроизводимости, учитывающая вариацию внутри серии наблюдений ($q = 1,2,...,n$).

Как и при однофакторном анализе, предполагается, что ε_{ijq} имеет нормальное распределение с нулевым математическим ожиданием и дисперсией $\sigma^2_{ош}$.

Если предположить, что между X_1 и X_2 взаимодействие отсутствует (что, строго говоря, требует специального исследования), то для результата каждого наблюдения можно представить моделью:

$$y_{ij} = \mu + \alpha_i + \beta_j + \varepsilon_{ij}.$$

Такая модель лежит в основе обработки результатов наиболее простого варианта - двухфакторного дисперсионного анализа без повторений. Таблица исходных данных для такого случая имеет следующий вид.

Таблица 6.3

Данные для двухфакторного дисперсионного анализа без повторений

Уровни фактора X_2	Уровни фактора X_1				Итоги
	x_{11}	x_{12}	…	x_{1k}	
x_{21}	y_{11}	y_{21}	…	y_{k1}	B_1
x_{22}	y_{12}	y_{22}	…	y_{k2}	B_2
…	…	…	…	…	…
x_{2m}	y_{1m}	y_{2m}	…	y_{km}	B_k
Итоги	A_1	A_2	…	A_k	

Расчеты рекомендуют выполнять по следующему алгоритму. Сначала, непосредственно в таблице исходных данных, следует вычислить итоги по столбцам и строкам:

$$A_i = \sum_{j=1}^{m} y_{ij} \text{ и } B_j = \sum_{i=1}^{k} y_{ij}.$$

Затем рассчитываются:

сумма квадратов всех наблюдений

$$SS_1 = \sum_{i=1}^{k} \sum_{j=1}^{m} y_{ij}^2 ;$$

средняя сумма квадратов итогов по столбцам

$$SS_2 = \frac{1}{m} \sum_{i=1}^{k} A_i^2 ;$$

средняя сумма квадратов итогов по строкам

$$SS_3 = \frac{1}{k} \sum_{j=1}^{m} B_j^2 ;$$

усредненный по всем наблюдениям квадрат общего итога

$$SS_4 = \frac{1}{mk} \left(\sum_{i=1}^{k} A_i \right)^2 = \frac{1}{mk} \left(\sum_{j=1}^{m} B_j \right)^2 ;$$

сумма квадратов для столбца

$$SS_A = SS_2 - SS_4 ;$$

сумма квадратов для строки

$$SS_B = SS_3 - SS_4 ;$$

общая сумма квадратов

$$SS_{общ} = SS_1 - SS_4 ;$$

остаточная сумма квадратов

$$SS_{ост} = SS_{общ} - SS_A - SS_B = SS_1 - SS_2 - SS_3 + SS_4 ;$$

дисперсия отклика от фактора X_1

$$s_{X1}^2 = \frac{SS_A}{k-1} ;$$

дисперсия отклика от фактора X_2

$$s_{X2}^2 = \frac{SS_B}{m-1} ;$$

дисперсия отклика от случайных причин

$$s_{ош}^2 = \frac{SS_{ост}}{(k-1)(m-1)} .$$

Результаты двухфакторного дисперсионного анализа рекомендуется представлять в виде таблицы 6.4.

Таблица 6.4

Результаты двухфакторного дисперсионного анализа без повторения опытов

Источник дисперсии	Число степеней свободы	Сумма квадратов	Средний квадрат	Математическое ожидание среднего квадрата
X_1	$k-1$	SS_A	s^2_{X1}	$m\sigma^2_{X1}+\sigma^2_{ош}$
X_2	$m-1$	SS_B	s^2_{X2}	$n\sigma^2_{X2}+\sigma^2_{ош}$
Остаток	$(k-1)(m-1)$	$SS_{ост}$	$s^2_{ош}$	$\sigma^2_{ош}$
Общая сумма	$km-1$	$SS_{общ}$	-	-

Влияние факторов следует признать значимым при выполнении условий:

$$F=\frac{s^2_{X1}}{s^2_{ош}}>F[\alpha;k-1;(k-1)(m-1)];$$

$$F=\frac{s^2_{X2}}{s^2_{ош}}>F[\alpha;m-1;(k-1)(m-1)].$$

Степени влияния факторов можно найти из выражений для математических ожиданий среднего квадрата:

$$\sigma^2_{X1}\approx\frac{s^2_{X1}-s^2_{ош}}{m}\ \text{и}\ \sigma^2_{X2}\approx\frac{s^2_{X2}-s^2_{ош}}{k}.$$

6.3.Пример дисперсионного анализа в *MS Excel* [1]

Проведение двухфакторного дисперсионного анализа без повторений рассмотрим на примере зависимости температуры переднего конца полосы на выходе из чистовой группы клетей стана от следующих факторов:

[1] Пример подготовлен доцентом, канд. техн. наук Б.Я. Омельченко

- Vмк=0÷300 м^3/ч – количество воды, подаваемой на полосу для ее охлаждения в межклетевых промежутках (фактор А);
- Vз=5÷12 м/с – скорость заправки полосы в моталки (фактор В).

Таблица 6.5

Влияние количества воды при межклетевом охлаждении и заправочной скорости на температуру переднего конца полосы, °С

	Vмк, м^3/ч					
Vз, м/с	*0*	*50*	*100*	*150*	*200*	*300*
5	752	797	741	724	731	749
6	834	788	792	783	744	771
7	826	800	824	793	818	827
8	825	807	855	840	812	780
9	889	892	853	868	806	845
10	911	882	887	886	894	887
11	927	875	925	871	879	842
12	940	904	949	934	956	898

Оформление листа рабочей книги показано на рис. 6.1. Исходные данные записаны в ячейки A5:G13. В ячейке H5 подсчитано количество вариантов фактора А по формуле =СЧЁТ(B5:G5). В ячейке A14 подсчитано количество вариантов фактора В по формуле =СЧЁТ(A6:A13). В ячейках H6:H13 определены средние значения для строк по формулам вида =СРЗНАЧ(B6:G6). В ячейках B14:G14 определены средние значения для столбцов по формулам вида =СРЗНАЧ(B6:B13). В ячейке H14 определено среднее арифметическое для всех наблюдений по формуле =СРЗНАЧ(B6:G13).

Ниже на листе выполнена вспомогательная таблица, в которой по формулам вида =(B6-H14)^2 найдены квадраты разностей:

- между отдельными значениями наблюдений и общим средним (ячейки B16:G23);
- между средними по фактору А и общим средним (ячейки H16:H23);
- между средними по фактору В и общим средним (ячейки B24:G24).

	A	B	C	D	E	F	G	H
1	Двухфакторный дисперсионный анализ без повторений							
2	Влияние на температуру переднего конца полосы (Тпк)							
3	расхода охлажд.воды (Vмк) и заправочной скорости (Vз)							
4								
5	Vз / Vмк	0	50	100	150	200	300	6
6	5	752	797	741	724	731	749	*749,00*
7	6	834	788	792	783	744	771	*785,33*
8	7	826	800	824	793	818	827	*814,67*
9	8	825	807	855	840	812	780	*819,83*
10	9	889	892	853	868	806	845	*858,83*
11	10	911	882	887	886	894	887	*891,17*
12	11	927	875	925	871	879	842	*886,50*
13	12	940	904	949	934	956	898	*930,17*
14	8	*863,00*	*843,13*	*853,25*	*837,38*	*830,00*	*824,88*	**841,938**
15								
16		8088,75	2019,38	10188,38	13909,25	12307,13	8637,38	8637,38
17		63,00	2909,25	2493,75	3473,63	9591,75	5032,13	3204,03
18		254,00	1758,75	321,75	2394,88	573,00	223,13	743,70
19		286,88	1220,63	170,63	3,75	896,25	3836,25	488,59
20		2214,88	2506,25	122,38	679,25	1291,50	9,38	285,47
21		4769,63	1605,00	2030,63	1941,50	2710,50	2030,63	2423,51
22		7235,63	1093,13	6899,38	844,63	1373,63	0,00	1985,82
23		9616,25	3851,75	11462,38	8475,50	13010,25	3143,00	7784,39
24		443,63	1,41	127,97	20,82	142,50	291,13	
25								
26	*Величина*	*S*	*k*	*Дисп*	*F* расч	*F* табл		
27	Фактор В (Vз)	153317,31	7	21902,47	42,508	2,285		
28	Фактор А (Vмк)	8219,69	5	1643,94	3,191	2,485		
29	Погрешность	18033,81	35	515,25				
30	Всего	179570,81	47					

Рис. 6.1. Пример двухфакторного дисперсионного анализа без повторений

В ячейках A26:F30 приведены сводные данные двухфакторного дисперсионного анализа без повторений:

S – сумма квадратов разностей;

k – число степеней свободы;

Дисп – дисперсия;

F расч – расчетное значение критерия Фишера;

F табл – табличное значение критерия Фишера.

Суммы квадратов разностей (*S*) рассчитаны по формулам:

Величина	№ формулы	Ячейка	Формула в *MS Excel*
Фактор В (Vз)	4.48	B27	=H5*СУММ(H16:H23)
Фактор А (Vмк)	4.42	B28	=A14*СУММ(B24:G24)
Погрешность	4.47	B29	=B30-B27-B28
Всего	4.40	B30	=СУММ(B16:G23)

Число степеней свободы (*k*) определены по формулам:

Величина	№ формулы	Ячейка	Формула в *MS Excel*
Фактор В (Vз)	4.52	C27	=A14-1
Фактор А (Vмк)	4.52	C28	=H5-1
Погрешность	4.52	C29	=(A14-1)*(H5-1)
Всего	4.52	C30	=A14*H5-1

Дисперсии (*Дисп*) найдены по формулам:

Величина	№ формулы	Ячейка	Формула в *MS Excel*
Фактор В (Vз)	4.51	D27	=B27/C27
Фактор А (Vмк)	4.51	D28	=B28/C28
Погрешность	4.50	D29	=B29/C29
Всего			

Расчетные значения критерия Фишера (*F* расч) определены по формулам:

Величина	№ формулы	Ячейка	Формула в *MS Excel*
Фактор В (Vз)	4.3	E27	=D27/D29
Фактор А (Vмк)	4.3	E28	=D28/D29
Погрешность			
Всего			

Табличные значения критерия Фишера (*F* табл) найдены для вероятности 0,05 по формулам:

Величина	№ формулы	Ячейка	Формула в *MS Excel*
Фактор В (Vз)	–	F27	=FРАСПОБР(0,05;C27;C29)
Фактор А (Vмк)	–	F28	=FРАСПОБР(0,05;C28;C29)
Погрешность			
Всего			

6.4.Обсуждение результатов

Сравнивая расчетные и табличные значения критерия Фишера можно сделать следующие выводы:

1. С надежностью 95% оба фактора и заправочная скорость, и количество охлаждающей воды являются значимыми, так как F расч > F табл в обоих случаях.

2. Заправочная скорость оказывает большее влияние на температуру полосы, чем количество воды для межклетевого охлаждения, так как разность F расч - F табл в данном случае больше, чем для Vмк.

Рассмотренный выше пример проведения дисперсионного анализа можно выполнить также, используя соответствующий инструмент по дисперсионному анализу из встроенного в *MS Excel* пакета ***Анализ данных***. В приложении 3 приведено описание работы с инструментом «Двухфакторный дисперсионный анализ без повторений».

6.5.Контрольные вопросы

1. Поясните сущность дисперсионного анализа и перечислите его основные допущения.

2. Поясните постановку задачи и запишите модель однофакторного дисперсионного анализа.

3. Запишите и поясните условие значимости влияния фактора на отклик. Как определить степень влияния фактора на отклик при однофакторном анализе с равным числом повторений?

4. Поясните постановку задачи и запишите модель двухфакторного дисперсионного анализа.

5. Запишите и поясните условие значимости влияния факторов на отклик для двухфакторного анализа. Как определить степень влияния фактора на отклик при двухфакторном анализе?