3 ОБРАБОТКА ВЫБОРКИ И РАСЧЕТ ОПИСАТЕЛЬНЫХ СТАТИСТИК

Изучение закономерностей всякой случайной величины возможно только по выборке значений из ее генеральной совокупности. Поэтому для получения достоверных значений параметров закона распределения, прежде всего, необходимо, чтобы выборка не содержала грубых погрешностей (значений, которые для данной случайной величины не характерны).

Если грубые погрешности в выборке отсутствуют, то достоверность найденных значений параметров распределения будет определяться особенностями соответствующих выборочных оценок. Наилучшие результаты обеспечиваются, если оценки обладают свойствами состоятельности, несмещенности и эффективности.

Таким образом, обработка и анализ выборки предусматривает выявление и устранении грубых погрешностей, расчет требуемых числовых характеристик с использованием обоснованных выборочных оценок (описательных статистик), а также выявление типа распределения анализируемой величины на основании выборочных данных.

Среди числовых характеристик случайной величины различают:

- характеристики положения (центральной тенденции);
- характеристики рассеяния (вариации);
- характеристики формы распределения (асимметрия и эксцесс).

3.1. Характеристики положения

Математическое ожидание. Математическое ожидание случайной величины X представляет собой такое ее значение M_x, около которого сосредоточены все другие возможные. Наилучшей оценкой математического ожидания (т. е. и состоятельной, и несмещенной, и эффективной) является выборочное среднее, рассчитываемое как среднее арифметическое:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \approx M_x . \qquad (3.1)$$

В *MS Excel* для расчета $\overline{x}$ имеется *статистическая функция* СРЗНАЧ(*Аргумент*). Аргументом функции является ссылка на ячейки

рабочего листа, в которых расположены элементы анализируемой выборки.

Медиана. Медиана Me – это такое значение случайной величины, что для 50% ее возможных значений выполняется условие $x < Me$, а для других 50% выполняется условие $x > Me$. На графике функции распределения (рис. 3.1,а) медиана есть абсцисса, которой соответствует значение $F(x)$=0,5. На графике плотности распределения (рис. 3.1,б) медиана есть абсцисса, которая делит площадь под кривой $f(x)$ на две равные части.

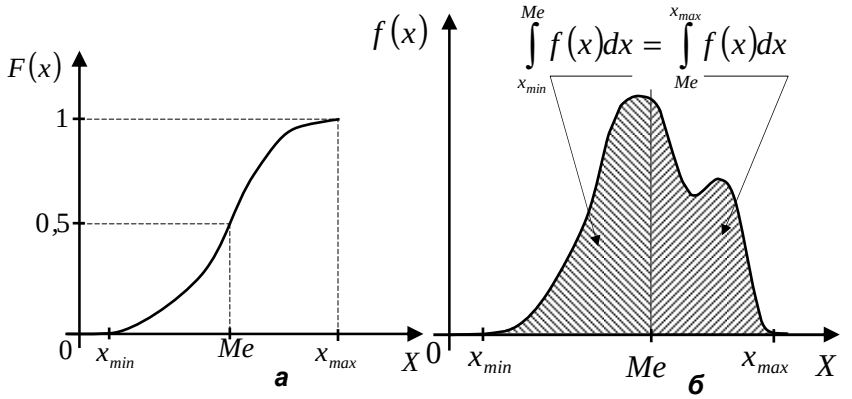

Рис. 3.1. Положение медианы:
а – на графике функции распределения;
б – на графике плотности распределения

Для нахождения медианы по выборке необходимо сначала выполнить ее ранжирование (расположить ее элементы) по возрастанию. Если объем выборки есть нечетное число, то оценкой медианы будет элемент, расположенный в середине ряда:

$$Me \approx x_{(n+1)/2}. \qquad (3.2)$$

Если объем выборки есть четное число, то:

$$Me \approx \frac{x_{n/2} + x_{n/2+1}}{2}. \qquad (3.3)$$

В *MS Excel* для расчета Me имеется статистическая функция МЕДИАНА().

Мода. Мода Mo – значение случайной величины, вероятность появления которого наибольшая. На графике плотности вероятности мода есть абсцисса, соответствующая максимуму кривой $f(x)$. При оценке моды по выборке Mo принимают равным тому значению случайной величины, которое встречается в выборке наиболее часто.

В *MS Excel* для расчета Me имеется статистическая функция МОДА(). Особенностью ее работы является то, что если в анализируемом наборе данных отсутствуют повторяющиеся значения, функция выдает сообщение:

#НЕТ ДАННЫХ#.

Для нормального распределения среднее выборочное, медиана и мода совпадают:

$$\bar{x} = Me = Mo\,.$$

3.2. Характеристики рассеяния (вариации)

Дисперсия. Дисперсия D_X – математическое ожидание квадратов отклонений значений случайной величины от ее математического ожидания:

$$D_X = M\left\{\left[x - M_X\right]^2\right\}. \qquad (3.4)$$

На основании выборки дисперсию случайной величины оценивают следующими характеристиками: *дисперсией распределения* σ^2 и *выборочной дисперсией* s^2. Указанные выборочные оценки дисперсии рассчитывают по формулам:

$$\sigma^2 = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2; \qquad (3.5)$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2. \qquad (3.6)$$

Выборочная дисперсия (3.6) является эффективной, несмещенной и состоятельной оценкой при любом объеме выборки. Для дисперсии распределения (3.5) состоятельность также обеспечивается при любом объеме выборки, но несмещенность и эффективность достигаются только при $n > 30$.

Несмещенность и эффективность s^2 достигнуты за счет того, что в знаменателе объем выборки уменьшен на единицу. Это оказалось необ-

ходимым в связи с использованием в формуле среднего выборочного $\bar{x}$, значение которого связано с элементами рассматриваемой выборки. Каждая величина, зависящая от элементов выборки и используемая в формуле выборочной оценки, называется *связью*. Разность между объемом выборки и числом связей l в формуле, по которой рассчитывается статистика, называют *числом степеней свободы* данной статистики $\nu = n - l$.

Для вычисления дисперсии распределения в *MS Excel* предусмотрена статистическая функция ДИСПР(), а для вычисления выборочной дисперсии – функция ДИСП().

Среднее квадратическое отклонение (стандартное отклонение, стандарт). Недостатком дисперсии считают то, что она имеет размерность квадрата анализируемой величины. Для устранения указанного недостатка было введено среднее квадратическое отклонение (стандартное отклонение, стандарт), которое представляет собой корень квадратный из дисперсии случайной величины:

$$\sigma_X = \sqrt{D_X} \ . \tag{3.7}$$

Наилучшей выборочной оценкой для σ_X является *выборочное среднее квадратическое отклонение (выборочное стандартное отклонение)*:

$$s = \sqrt{s^2} = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \approx \sigma_X \ . \tag{3.8}$$

Как видно из формулы (3.8), выборочное стандартное отклонение определяется через выборочную дисперсию (3.6) и поэтому также является несмещенной и эффективной оценкой при любом *n*. В *MS Excel* для расчета *s* предусмотрена статистическая функция СТАНДОТКЛОН().

Для выборок объемом $n > 30$ свойства несмещенности и эффективности проявляются также у стандартного отклонения, вычисляемого на основании дисперсии распределения (3.5):

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}\right)^2} \approx \sigma_X \ . \tag{3.9}$$

В *MS Excel* параметр σ рассматривается как *стандартное отклонение генеральной совокупности* и вычисляется с помощью статистической функции СТАНДОТКЛОНП(). В литературе можно также встретить термин «*генеральный стандарт*»

В практике анализа числовой информации используют также и ряд других характеристик рассеяния случайной величины.

Размах. Размах R – разность между наибольшим и наименьшим значениями случайной величины, обнаруженными в выборке:

$$R = x_{max} - x_{min} . \qquad (3.10)$$

Этот показатель не дает представления об особенностях рассеяния случайной величины и может существенно меняться при переходе от одной выборки из генеральной совокупности к другой выборке из той же совокупности.

Коэффициент вариации. Коэффициент вариации V_x характеризует, какую долю от математического ожидания случайной величины составляет ее среднее квадратическое отклонение:

$$V_x = \frac{s}{\bar{x}}. \qquad (3.11)$$

Будучи безразмерной величиной, коэффициент вариации позволяет сравнивать степень рассеяния различных случайных величин. Его значение может быть представлено как в относительных единицах, так и в процентах. В последнем случае результат, полученный расчетом по формуле (3.9), необходимо умножить на 100. Недостатком этого показателя является то, что он становится малонадежным при $\bar{x} \to 0$, а при $\bar{x}$ =0 вообще теряет смысл.

3.3. Характеристики формы распределения

Коэффициент асимметрии. Асимметрия распределения проявляется в том, что значения случайной величины $x_1 = M_x - \Delta$ и $x_2 = M_x + \Delta$ (т. е. удаленные в разные стороны от математического ожидания на одну и ту же величину Δ) имеют различные вероятности. В связи с этим при наличии асимметрии мода и медиана не совпадают (рис. 3.2).

Для характеристики асимметрии распределения применяют *коэффициент асимметрии*, выборочная оценка которого имеет вид:

$$A = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^3 . \qquad (3.12)$$

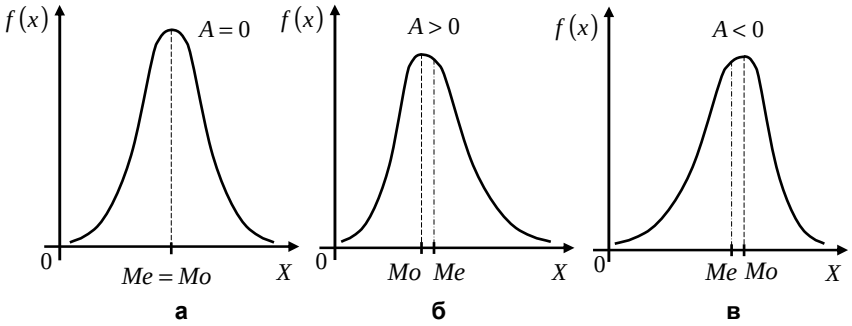

Рис. 3.2. Асимметрия распределения

При $A > 0$ (положительная асимметрия) вытянута правая ветвь плотности распределения, при $A < 0$ (отрицательная асимметрия) вытянута левая ветвь распределения. Для нормального и любого другого симметричного распределения A=0.

Асимметрию распределения называют также *скошенностью*. По-видимому, именно в связи с этим для расчета коэффициента асимметрии в *MS Excel* используется статистическая функция СКОС().

С применением коэффициента асимметрии можно выполнить приблизительную оценку соответствия выборочного распределения нормальному закону. Условие соответствия:

$$|A|/s_A < 3 \ . \qquad (3.13)$$

где s_A - стандартное отклонение асимметрии:

$$s_A = 3\sqrt{\frac{6n(n-1)}{(n-2)(n+1)(n+3)}} \ . \qquad (3.14)$$

Коэффициент эксцесса. Характеризует остроконечность или сглаженность распределения изучаемой случайной величины по сравнению с нормальным законом (рис. 3.3). Положительный эксцесс ($E > 0$) означает, что распределение изучаемой случайной величины более остроконечное, чем нормальное. Отрицательный эксцесс ($E < 0$) указывает на

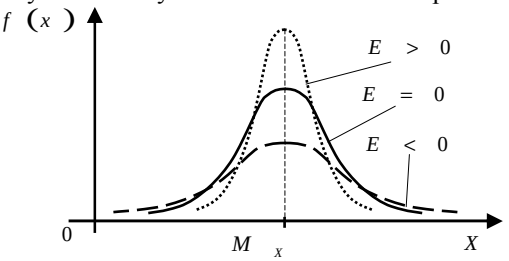

Рис. 3.3. Эксцесс распределения

сглаженность изучаемого распределения по сравнению с нормальным. Выборочная характеристика эксцесса имеет вид:

$$E \approx \left[\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum_{i=1}^{n} \left(\frac{x_i - \bar{x}}{s} \right)^4 \right] - \frac{3(n-1)}{(n-2)(n-3)}. \quad (3.15)$$

Для расчета E в *MS Excel* предусмотрена статистическая функция ЭКСЦЕСС().

С применением коэффициента эксцесса можно выполнить приблизительную оценку соответствия выборочного распределения нормальному закону. Условие соответствия:

$$|E| / s_E < 3. \quad (3.16)$$

где s_E - стандартное отклонение эксцесса:

$$s_E = (n-1) \sqrt{\frac{24n}{(n-3)(n-2)(n+3)(n+5)}}. \quad (3.17)$$

3.4. Интервальные оценки

Интервальные оценки характеризуют ошибку оценивания истинного значения ξ некоторой характеристики распределения случайной величины с помощью соответствующей выборочной оценки θ. Их применение необходимо потому, что каждая выборочная оценка сама по себе является случайной величиной с некоторым распределением вероятности. При интервальном оценивании используют *доверительную вероятность*, *уровень значимости*, *доверительный интервал* и *доверительные границы*.

Доверительная вероятность – вероятность события, заключающегося в том, что ошибка оценивания истинного значения некоторого параметра распределения случайной величины его выборочной оценкой не превышает величины Δ:

$$p = Prob\left(|\xi - \theta| \le \Delta \right). \quad (3.18)$$

Уровень значимости – вероятность события, заключающегося в том, что ошибка оценивания истинного значения некоторого параметра распределения случайной величины его выборочной оценкой превышает величину Δ:

$$\alpha = Prob\left(|\xi - \theta| > \Delta \right). \quad (3.19)$$

Доверительная вероятность и уровень значимости связаны соотношением:

$$p + \alpha = 1. \tag{3.20}$$

Доверительный интервал – интервал значений выборочной характеристики, внутри которого истинное значение оцениваемого параметра находится с заданной доверительной вероятностью.

Доверительные границы – значения выборочной оценки, представляющие собой границы доверительного интервала:

$$\theta_1 = \theta - \Delta; \tag{3.21}$$

$$\theta_2 = \theta + \Delta. \tag{3.22}$$

В практике обработки и анализа числовой информации наиболее часто встречается задача интервального оценивания выборочного среднего $\overline{x}$. При ее решении исходят из того, что $\overline{x}$ как случайная величина имеет нормальное распределение с математическим ожиданием $M_{\overline{x}} = \overline{x}$ и средним квадратическим отклонением

$$s_{\overline{x}} = \frac{s}{\sqrt{n}}. \tag{3.23}$$

Величина $s_{\overline{x}}$ имеет также специальное название - *стандартная ошибка выборочного среднего*.

Доверительные границы для выборочного среднего симметричны:

$$\Delta_x = \pm\, s_{\overline{x}}\, t[\alpha; n-1], \tag{3.24}$$

где $t[\alpha; n-1]$ – табличное значение (квантиль) распределения Стьюдента при уровне значимости α и числе степеней свободы $\nu = n - 1$.

Существуют специальные справочные таблицы распределения Стьюдента, которые, как правило, обязательно приводятся в литературе по вопросам теории вероятности и математической статистике. В *MS Excel* табличное число Стьюдента легко определить с помощью статистической функции СТЬЮДРАСПОБР(α, ν).

Зная доверительные границы выборочного среднего, можно утверждать, что с вероятностью $p = 1 - \alpha$ истинное значение случайной величины равно

$$\overline{x} \pm \Delta_{\overline{x}}. \tag{3.25}$$

Иначе:

$$\overline{x} - \Delta_{\overline{x}} \le \mu \le \overline{x} + \Delta_{\overline{x}}. \tag{3.26}$$

3.5. Отсеивание грубых погрешностей (проверка однородности выборки)

Выборка из генеральной совокупности получается в результате использования тех или иных средств измерения. Вследствие поломки прибора или по недосмотру исследователя среди элементов выборки могут оказаться результаты, содержащие грубую ошибку (погрешность). Ошибочные значения в дальнейшем анализе искажают информацию об исследуемом явлении. Поэтому перед тем, как использовать выборочные данные для каких-либо выводов, необходимо исследовать их на наличие грубых погрешностей.

Для анализа закономерностей случайной величины по выборке необходимо также иметь гарантию, что полученная выборка сделана из генеральной совокупности именно того параметра, который интересует исследователя, и в ней отсутствуют элементы из какой-либо иной генеральной совокупности. Выборку, которая содержит элементы только одной генеральной совокупности, принято называть *однородной*, а выявление «чужеродных» элементов – *проверкой однородности*.

И обнаружение грубых погрешностей, и проверка однородности по существу сводятся к одной и той же задаче – среди элементов выборки необходимо выявить такие, которые не характерны для исследуемой генеральной совокупности. Известно достаточно много методов решения этой задачи. Рассмотрим метод, рекомендуемый Е. Н. Львовским.

Сначала необходимо определить среднее выборочное $\bar{x}$ и несмещенное среднее квадратическое отклонение $\bar{s}$.

Затем среди элементов выборки необходимо выявить результат, который дает наибольшее по абсолютной величине отклонение от среднего выборочного:

$$d_{max} = max\left\{\left|x_i - \bar{x}\right|\right\}. \tag{3.27}$$

Далее вычисляются статистики τ_{max} и $\tau[\alpha; n-2]$ для уровней значимости 5 и 0,1%:

$$\tau_{max} = \frac{d_{max}}{\bar{s}}; \tag{3.28}$$

$$\tau[\alpha; n-2] = \frac{t[\alpha; n-2]\sqrt{n-1}}{\sqrt{(t[\alpha; n-2])^2 + (n-2)}}, \tag{3.29}$$

где $t[\alpha; n-2]$ – табличное число Стьюдента для уровня значимости α и числа степеней свободы $\nu = n-2$.

Значения статистик сравниваются, и делается вывод:

$\tau_{max} \le \tau[0{,}05; n-2]$ – элемент выборки не является аномальным;

$\tau[0{,}05; n-2] < \tau_{max} \le \tau[0{,}001; n-2]$ – возможно, что элемент выборки является аномальным;

$\tau_{max} > \tau[0{,}001; n-2]$ – элемент выборки является аномальным.

При обнаружении аномального элемента его необходимо исключить из выборки и повторить все предыдущие действия. Если же в отношении аномальности элемента выборки есть сомнения (второй случай соотношения статистики τ_{max} и $\tau[\alpha; n-2]$), то его можно оставить или отсеять по усмотрению исследователя.

3.6. Пример обработки выборки в MS EXCEL

Пример оформления листа приведен на рис. 3.4. Здесь представлены результаты анализа и обработки выборки, полученной по результатам наблюдений за толщиной полос, прокатываемых на широкополосном стане горячей прокатки при его настройке на номинал 2,5 мм.

Исходные данные (ячейки B2:B27) вводятся с клавиатуры.

3.6.1. Выявление и отсеивание грубых погрешностей в выборке

Сначала в ячейках G3, G4 и G5 с использованием статистических функций: СЧЕТ(B3:B27), СРЗНАЧ(B3:B27) и СТАНДОТКЛОН(B3:B27) программируется вычисление объема выборки (на рабочем листе обозначено как n), среднего выборочного ($X_{\text{ср}}$) и выборочного среднеквадратического отклонения (S).

Внимание! Здесь и далее аргументы статистических функций и адреса отдельных ячеек записаны в соответствии с расположением данных в примере, приведенном на рис.3.4.

Затем в ячейках D3:D27 программируется расчет отклонения текущего значения случайной величины от среднего выборочного:

$$|d_i| = |x_i - \bar{x}|.$$

Например, в ячейке D3:

=ABS(B3-$G$4).

После этого в ячейку G6 вводится формула, обеспечивающая выбор наибольшего из отклонений:

=МАКС(D3:D27).

Далее в ячейки G7:G11 вводятся формулы для получения данных, которые позволят оценить отклонение dmax. Для рассматриваемого примера:

Ячейка	*Формула*
G7	=G6/G5
G8	=СТЬЮДРАСПОБР(0,05;G3-2)
G9	=СТЬЮДРАСПОБР(0,001;G3-2)
G10	=G8*КОРЕНЬ(G3-1)/(КОРЕНЬ(G8^2+(G3-2)))
G11	=G9*КОРЕНЬ(G3-1)/(КОРЕНЬ(G9^2+(G3-2)))

В ячейку F14 необходимо запрограммировать автоматический вывод результата оценивания $d_{\max}$. С этой целью можно использовать логическую функцию ЕСЛИ(). Для рассматриваемого примера:

=ЕСЛИ(G7<G10;”ОШИБКИ НЕТ”;
ЕСЛИ(G7>G11;”ОШИБКА!”;”МОЖЕТ БЫТЬ”))

При получении результатов «ОШИБКА!» или «МОЖЕТ БЫТЬ» значения x_i и $\left| d_i \right|$, расположенные в строке, в которой |di |= dmax, необходимо удалить. Процедура повторяется до удаления всех выявленных элементов.

3.6.2. Определение выборочных характеристик

После обеспечения однородности выборки вычислить выборочные характеристики с применением инструмента «Описательная статистика» (описание работы с инструментом приведено в приложениях 1 и 2). В примере на рис. 3.4 результаты работы инструмента представлены в ячейках I3:J18. Используя клавиатуру, их следует дополнить размерностями каждой из выборочных характеристик (ячейки K5:K18).

Затем вычислить каждую из характеристик с использованием функций *MS Excel* и полученные значения отформатировать с использованием формата ячейки типа «Числовой». В примере соответствующие значения представлены в ячейках L5:L18.

3.6.3. Оценка нормальности распределения

С использованием описательных статистик нормальность распределения удобно оценивать по асимметрии и эксцессу.

Сначала в J21 рассчитывается стандартное отклонение асимметрии S_A:

=КОРЕНЬ(6*$J$17*($J$17-1)/($J$17/2)/($J$17+1)/($J$17+3)).

Затем в ячейке J22 вычисляется отношение выборочной асимметрии к ее стандартному отклонению:

=ABS(J12)/(J21).

Вывод по результатам сравнения в ячейке J23 формируется с помощью функции ЕСЛИ:

=ЕСЛИ(J22<3;"Да";"Нет").

Затем в ячейке J24 рассчитывается стандартное отклонение эксцесса S_E:

=($J$17-1)*КОРЕНЬ(24*$J$17/($J$17-3)/($J$17-2)/($J$17+3)/($J$17+5)),

в ячейке J25 определяется отношение $|E|/S_E$:

=ABS(J11)/(J24),

а в ячейке J26 формируется вывод относительно эксцесса:

=ЕСЛИ(J25<3;"Да";"Нет").

В обоих случаях вывод «Да» означает, что распределение вероятности анализируемого параметра можно считать нормальным.

На основании полученных результатов можно сделать следующие выводы.

1. При проверке однородности выборки был исключен один элемент, так как он являлся грубой погрешностью.

2. Истинное значение случайной величины равно 2,5 ± 0,01 мм, т.е. находится в пределах от 2,49 до 2,51 мм.

	A	B	C	D	E	F	G	H	I	J	K	L
1		**ВЫБОРКА**		**ОТКЛОНЕНИЕ**		**ПРОВЕРКА**			**ВЫБОРОЧНЫЕ ХАРАКТЕРИСТИКИ**			
2		h, мм		|di|,мм		**ОДНОРОДНОСТИ**			**"ОПИСАТЕЛЬНАЯ СТАТИСТИКА"**			**Функции**
3		2,50		0,00		n	24		*h, мм*			***Excel***
4		2,47		0,03		Xcp	2,498					
5		2,51		0,01		S	0,025		Среднее	2,4975	мм	2,50
6		2,50		0,00		dmax	0,05		Стандартная ошибка	0,005187	мм	0,005
7		2,45		0,05		τ_{max}	1,869		Медиана	2,5	мм	2,50
8		2,48		0,02		t [0,05;n-2]	2,074		Мода	2,5	мм	2,50
9		2,52		0,02		t [0,001;n-2]	3,792		Стандартное отклонение	0,02541	мм	0,025
10		2,49		0,01		τ [0,05;n-2]	1,939		Дисперсия выборки	0,000646	мм2	0,0006
11		2,50		0,00		τ [0,001;n-2]	3,015		Эксцесс	-0,50664	-	-0,51
12		2,47		0,03					Асимметричность	0,045535	-	0,05
13		2,46		0,04		**ОШИБКИ НЕТ**			Интервал	0,09	мм	0,09
14		2,51		0,01					Минимум	2,45	мм	2,45
15		2,54		0,04					Максимум	2,54	мм	2,54
16		2,50		0,00					Сумма	59,94	мм	59,94
17		2,46		0,04					Счет	24	-	24
18		2,53		0,03					Уровень надежности(95,0%)	0,01073	мм	0,01
19									**ОЦЕНКА**			
20		2,54		0,04					**НОРМАЛЬНОСТИ РАСПРЕДЕЛЕНИЯ**			
21		2,49		0,01					**По асимметрии** S_A	0,639		
22		2,49		0,01					|A| / S_A	0,071		
23		2,48		0,02					Вывод	**Да**		
24		2,50		0,00					**По эксцессу** S_E	0,918		
25		2,54		0,04					|E| / S_E	0,552		
26		2,51		0,01					Вывод	**Да**		
27		2,50		0,00								
28												

Рис. 3.4.Пример оформления рабочего листа

3. В рассмотренном случае для распределения толщины характерна положительная асимметрия (A = 0,04553 > 0), однако ее величина не противоречит гипотезе о нормальности распределения анализируемой случайной величины.

4. Эксцесс распределения отрицательный ($E \approx -0{,}51$), что указывает на относительную сглаженность распределения данной случайной величины по сравнению с нормальным. Однако величина эксцесса не противоречит гипотезе о нормальности распределения анализируемой случайной величины.

3.7. Контрольные вопросы

1. Выборочный метод изучения случайной величины
2. Описательные статистики. Математическое ожидание и его выборочная оценка
3. Описательные статистики. Дисперсия и ее наилучшая выборочная оценка
4. Описательные статистики. Стандартное отклонение и его наилучшая выборочная оценка
5. Описательные статистики. Оценивание истинного значения параметра по выборке.
6. Однородность выборки и ее обеспечение.
7. Методы оценивания нормальности распределения. Сущность оценки нормальности по асимметрии и эксцессу.