Федеральное государственное бюджетное образовательное учреждение

высшего образования
«Ставропольский государственный аграрный университет»

Кафедра Математика

Составитель: доцент Литвин Д.Б.

МЕТОДИЧЕСКИЕ УКАЗАНИЯ ПО ВЫПОЛНЕНИЮ РГР №7

по дисциплине

МАТЕМАТИКА

наименование дисциплины

21.03.02 Землеустройство

направление подготовки

Городской кадастр

профиль(и) подготовки

Бакалавр

Квалификация (степень) выпускника

Ставрополь, 2019

ПЕРВИЧНАЯ ОБРАБОТКА РЕЗУЛЬТАТОВ НАБЛЮДЕНИЙ

1.1. Вариационный и статистический ряды

Пусть для изучения некоторого количественного признака X из всей объема N генеральной совокупности однородных объектов извлечена выборка объема n. Наблюдавшиеся значения $x_1, x_2, ..., x_k$ признака X называют *вариантами*, а последовательность вариант, записанных в возрастающем порядке, – *вариационным рядом*.

Статистическим распределением (рядом) выборки называют перечень вариант x_i вариационного ряда и соответствующих им частот n_i (сумма всех частот равна объему выборки n) или относительных частот $w_i = n_i / n$ (сумма всех относительных частот равна единице).

Различают *дискретный (точечный) и интервальный (сгруппированный)* статистический ряд.

Интервальный статистический ряд

Если имеется выборка значений *непрерывного* количественного признака, где число вариант очень велико, то составляется *сгруппированный (интервальный) статистический ряд*. Для его получения интервал (a, b), содержащий все варианты, делится на k равных частей длины h, и в качестве абсолютных частот выступают количества вариант, попавших в данный интервал.

Количество интервалов k следует выбирать так, чтобы построенный ряд не был громоздким, но в то же время позволял выявлять характерные изменения случайной величины.

Для вычисления k рекомендуется использовать формулу Стерджеса:

$$k = 1 + \log_2 n \quad \text{или} \quad k = 1 + \log_2 10 \cdot \lg n, \text{ где } \log_2 10 \approx 3{,}322 \qquad (1)$$

с округлением k до ближайшего целого значения.

Необходимо, чтобы интервал (a, b) статистического ряда длины kh с небольшим "нахлестом" перекрывал вариационный размах наблюдаемого признака

$$kh = b - a \geq x_{\max} - x_{\min}, \text{ т.е. чтобы } \quad a \leq x_{\min}, \qquad b \geq x_{\max}, \qquad (2)$$

поэтому длину частного интервала выбирают из неравенства

$$h \geq \frac{x_{\max} - x_{\min}}{k}, \qquad (3)$$

где $x_{\max}, x_{\min}$ – наибольшее и наименьшее значения признака.

После нахождения частных интервалов определяется, сколько значений случайной величины попало в каждый конкретный интервал

$$[x_i, x_i + h), \quad i = 1, 2, ..., k, \tag{4}$$

где k – число интервалов.

При этом в интервал включают значения, большие или равные нижней границе и меньшие верхней границы.

Для наглядного представления распределения наблюдаемого непрерывного признака X, исследуемого по выборке, используется ***гистограмма*** – столбчатая диаграмма, состоящая из прямоугольников, основания которых – частичные интервалы длины h, а высоты – плотности абсолютных n_i/h или относительных w_i/h частот (частостей). Гистограмма является статистическим аналогом плотности распределения *f(x),* при этом общая площадь гистограммы относительных частот (частостей) равна единице, а гистограммы абсолютных частот - объему выборки.

Эмпирическая функция распределения $F_n^*(x)$ для интервального статистического ряда является кусочно-линейной. При этом, она равна нулю в начале первого частного интервала, а в конце каждого частного интервала i принимает значения соответствующих накопленных частостей $w_i^{нак}$.

Пример 1.2. Отклонение диаметра вала после шлифовки от номинального значения в мм представлено следующей выборкой (объемом $n = 69$):

-0,84	0,26	0,88	-0,33	0,72	-0,44	0,93
-0,14	0,71	-0,99	1,21	0,31	-0,29	0,79
0,34	0,38	-0,54	1,81	1,13	0,28	1,22
0,79	-0,89	0,89	-0,49	-0,03	0,44	-0,35
-0,3	1,34	-0,8	-0,32	1,15	-0,41	0,76
0,25	-0,18	-0,41	0,96	-0,63	0,86	0,8
-0,61	-0,65	-0,03	1,72	1,96	0,45	-0,6
1,15	0,19	0,35	0,5	0,77	0,91	-0,26
0,51	1,36	-0,01	0,42	0,63	-0,14	0,1
-0,08	-0,97	0,55	0,38	0,86	-0,57	

Необходимо построить интервальный вариационный ряд, гистограмму относительных частот и эмпирическую функцию распределения (кумуляту).

Решение.

1. Определим количество частных интервалов по формуле (1):

$$k = 1 + \log_2 n = 1 + \log_2 69 = 7{,}108524.$$ Примем $k = 7$.

2. Выполним ранжирование вариант в порядке возрастания.
3. Определим длину частного интервала по формуле (3):

$$x_{\max} = 1{,}96; \quad x_{\min} = -0{,}99;$$

$$h \geq \frac{x_{\max} - x_{\min}}{k} = \frac{1{,}96 - (-0{,}99)}{7} = \frac{2{,}95}{7} = 0{,}421.$$ Примем $h = 0{,}43.$

4. Проверим выполнение условия (2):

$$kh \geq x_{\max} - x_{\min}; \quad 7 \cdot 0{,}43 \geq 1{,}96 - (-0{,}99); \quad 3{,}01 \geq 2{,}95.$$

При этом, "нахлест" диапазона статистического ряда составляет $3{,}01\text{-}2{,}95 = 0{,}06$. Это значение определяет максимальное смещение влево от $x_{\min}$ начала интервала (a, b) статистического ряда. В рассматриваемом примере смещение удобно принять *0,01*, т.е. левую границу примем

$$a = -1 \leq x_{\min} = -0{,}99,$$

тогда правая граница $b = a + kh = 2{,}01 \geq x_{\max} = 1{,}96$.

5. Определим границы частичных интервалов по формуле (4):

-1	-0,57	-0,14	0,29	0,72	1,15	1,58	2,01

6. Подсчитав количество вариант, попавших в соответствующие интервалы, получим искомый интервальный статистический ряд:

Таблица 1.2

i	1	2	3	4	5	6	7
X	[-1,00; -0,57)	[-0,57; -0,14)	[-0,14; 0,29)	[0,29; 0,72)	[0,72; 1,15)	[1,15; 1,58)	[1,58; 2,01)
n_i	9	13	11	13	14	6	3
$n_i^{нак}$	9	22	33	46	60	66	69
w_i	0,130	0,188	0,159	0,188	0,203	0,087	0,043
w_i / h	0,303	0,438	0,371	0,438	0,472	0,202	0,101
$w_i^{нак}$	0,130	0,319	0,478	0,667	0,870	0,957	1

7. Используя интервальный ряд (см. табл. 1), построим гистограмму плотностей относительных частот w_i/h, и эмпирическую функцию распределения $F_n^*(x)$, которые представлены на рисунке 1.2.

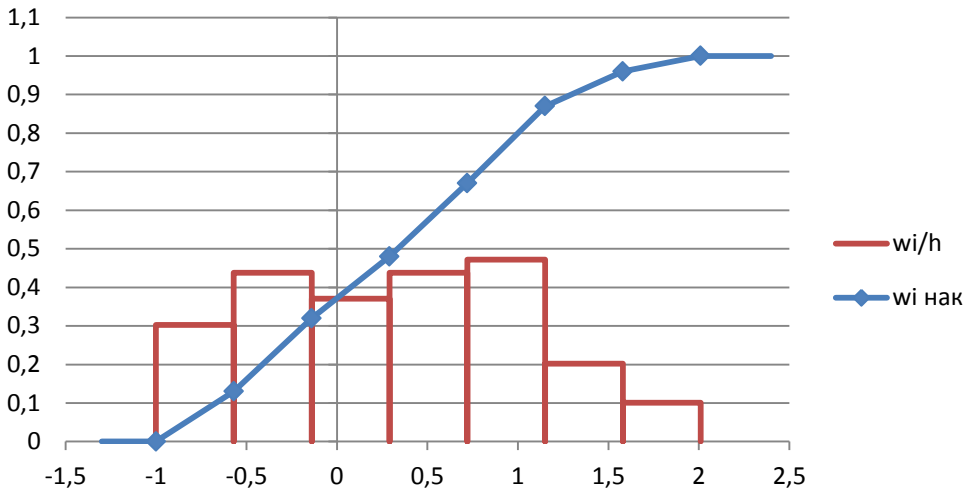

Рисунок 1.2 - Гистограмма и эмпирическая функция распределения

1.2. Первичная обработка в Excel

После внесения информации в электронную таблицу необходимо определить минимальный и максимальный варианты, размах вариационного ряда, количество вариант и число частных интервалов, например по формуле Стерджеса, и их границы. Затем вычислить частоты n_i. Эти операции можно сделать с помощью обычной сортировки. Существуют также встроенные специальные статистические функции.

Для вычисления частот $\mathbf{n_i}$ можно использовать ***функцию ЧАСТОТА***, обращение к которой имеет вид:

=ЧАСТОТА(*массив_данных;массив_интервалов*),

где *массив_данных* – адреса ячеек, для которых вычисляется частота n_i - количество вариант, меньших или равных граничному значению $(z_{i-1}; z_i]$, $i = 1,2,...,k+1$ (сравните с (4)); *массив_границ* – адреса ячеек, в которых размещаются упорядоченные по возрастанию значения z_i, $i = 1,2,...,k+1$, где k – число интервалов.

Особенности.

Количество элементов в возвращаемом массиве на единицу больше числа элементов в массиве "*массив_интервалов*". Дополнительный элемент в возвращаемом массиве содержит количество значений, превышающих верхнюю границу интервала, содержащего наибольшие значения.

Функция ЧАСТОТА вводится как формула массива, т.е. предварительно выделяется интервал ячеек, в который будут помещены вычисленные частоты (число ячеек должно быть на 1 больше числа границ), затем вводится функция ЧАСТОТА с соответствующими аргументами, потом одновременно нажимаются клавиши [Ctrl] + [Shift] + [Enter].

Функция МАКС вычисляет максимальное значение из заданных аргументов. Обращение к ней имеет вид:

=МАКС(*арг1; арг2; …; арг255*),

где *арг1; арг2; …; арг255* – числовые константы или адреса ячеек, содержащих числовые величины.

Функция МИН вычисляет минимальное значение из заданных аргументов. Обращение к ней имеет вид:

=МИН(*арг1; арг2; …; арг255*),

где *арг1; арг2; …; арг255* – числовые константы или адреса ячеек, содержащих числовые величины.

Для подсчета количества элементов выборки (т.е. объема выборки) использовалась ***функция СЧЁТ***, обращение к которой имеет вид:

СЧЁТ(*массив_данных*),

где *массив_данных* – адреса ячеек или числовые константы.

Построение гистограммы частот и частостей возможно как с использованием стандартных инструментов Excel (Вставка-Гистограмма), так и с помощью специального инструмента из вкладки *Данные-Анализ данных*, в которой следует выбрать пункт *Гистограмма*.

Появится окно гистограммы, показанное на рисунке 1.3.

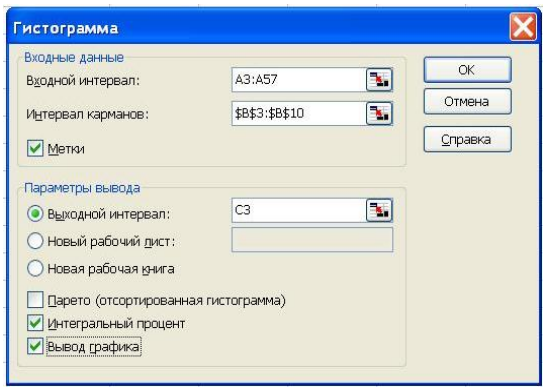

Рисунок 1.3 - Диалоговое окно режима ***Гистограмма***

В окне задаются следующие параметры:

Входной интервал: – адреса ячеек, содержащие выборочные данные.

Интервал карманов: (необязательный параметр) – адреса ячеек, содержащие границы интервалов (кармана). Эти значения должны быть введены в возрастающем порядке.

Если границы интервалов не заданы, то автоматически будет создан набор интервалов с одинаковой длиной

$$h = \frac{x_{\max} - x_{\min}}{[k]-1},$$

где $[k]$ – целая часть величины $k = 1 + 3{,}322 \cdot \lg n$, n – объем выборки.

Метки – флажок, включаемый, если первая строка во входных данных содержит заголовки. Если заголовки отсутствуют, то флажок следует выключить.

Парето (*отсортированная гистограмма*) – устанавливается в активное состояние, чтобы представить w_i в порядке их убывания. Если параметр выключен, то w_i приводятся в порядке следования интервалов.

Интегральный процент – устанавливается в активное состояние для расчета выраженных в процентах накопленных относительных частот (процентный аналог значений выборочной функции распределения.

Результатом использования описываемого инструмента для данных **примера 1.2** является гистограмма, представленная на рисунке 1.4 (сравните этот результат с таблицей 1.1 и рисунком 1.2).

Таблица 1.2

h	*Частота*	*Интегральный %*
-1	0	0,00%
-0,57	8	11,76%
-0,14	13	30,88%
0,29	11	47,06%
0,72	14	67,65%
1,15	15	89,71%
1,58	4	95,59%
2,01	3	100,00%
Еще	0	100,00%

Особенность.

Столбцы гистограммы и "Интегральный %" строятся по серединам частных интервалов.

Построенная гистограмма является ненормированной: высоты прямоугольников в ней равны частостям w_i, а не их плотностям w_i/h. В этом случае единице равна сумма высот всех прямоугольников, а не сумма их площадей. Поэтому ненормированная гистограмма не может служить оценкой для плотности распределения случайной величины, из значений которой была сформирована выборка (особенно в случае неравных длин интервалов) [5].

Поэтому для построения гистограммы и эмпирической функции распределения рекомендуется использовать стандартные инструменты Excel.

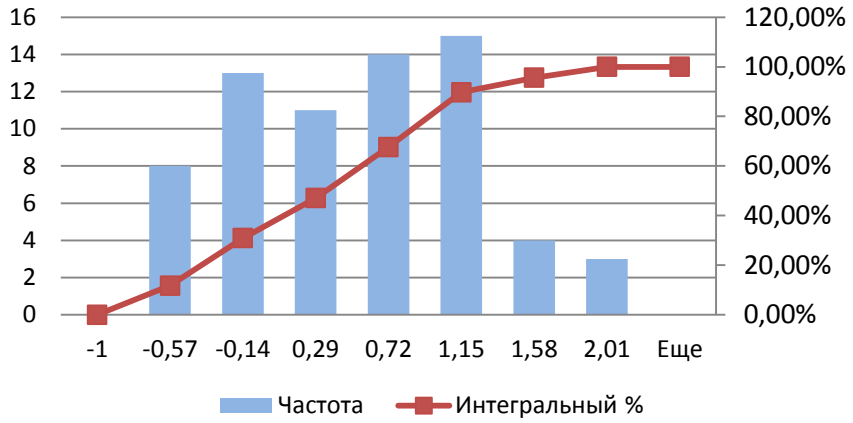

Рисунок 1.4 - График построенной гистограммы

2. ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

2.1. Требования к оценкам

Любая оценка (приближенное значение) параметра распределения вычисляется в статистике как функция случайных вариант наблюдаемого признака, а потому сама является в определенной мере случайной.

К статистическим оценкам обычно предъявляются требования:

- *состоятельности* - при увеличении числа наблюдений n она должна приближаться (сходиться по вероятности) к истинному значению параметра;

- *несмещенности* - ее математическое ожидание должно равняться значению параметра (отсутствие систематической ошибки);

- *эффективности* - среди всех несмещенных оценок выбранная должна обладать наименьшей дисперсией.

2.2. Точечные оценки M(X) и D(X)

По имеющейся выборке можно дать оценку математического ожидания и дисперсии генеральной совокупности. Несмещенной оценкой математического ожидания служит ***выборочное среднее***

$$\overline{x}_B = \frac{n_1 x_1 + n_2 x_2 + ... + n_k x_k}{n} = \frac{1}{n}\sum_i n_i x_i, \quad n = n_1 + n_2 + ... + n_k, \qquad (5)$$

то есть среднее арифметическое всех элементов выборки.

Оценкой дисперсии может служить ***выборочная дисперсия***

$$D_B = \sum_{i=1}^{k} \frac{n_i (x_i - \overline{x}_B)^2}{n} = \overline{\left(x - \overline{x}_B\right)^2}. \qquad (6)$$

Более удобна формула - *средний квадрат минус квадрат среднего*:

$$D_B = \sum_{i=1}^{k} \frac{n_i x_i^2}{n} - \overline{x}_B^2 = \overline{x^2} - \left(\overline{x}_B\right)^2. \qquad (7)$$

Выборочная дисперсия - смещенная в сторону занижения оценка генеральной дисперсии $D_Г$, и ее математическое ожидание $M(D_B) = \frac{n-1}{n} D_Г$. Поэтому вводится несмещенная оценка генеральной дисперсии – ***исправленная выборочная дисперсия***

$$s^2 = \frac{n}{n-1} D_B. \qquad (8)$$

Соответственно $s = \sqrt{s^2}$ является ***исправленным выборочным средним квадратическим отклонением.***

Замечание 1. Если первоначальные варианты x_i — большие числа, то для упрощения расчета целесообразно вычесть из каждой варианты одно и то же число C, т. е. перейти к *условным вариантам* $u_i = x_i - C$ (в качестве C выгодно принять число, близкое к выборочной средней; поскольку выборочная средняя неизвестна, число C выбирают «на глаз»).

Тогда для выборочного среднего справедливо

$$u_i = x_i - C \quad \rightarrow \quad \overline{x}_B = C + \frac{1}{n}\sum_i n_i u_i \ \text{ или } \ \overline{x}_B = C + \overline{(x - C)}; \qquad (9)$$

выборочная дисперсия при этом не изменится

$$u_i = x_i - C \quad \rightarrow \quad D_B(X) = D_B(u) = \overline{u^2} - \left(\overline{u}\right)^2 \qquad (10)$$

Замечание 2. Если первоначальные варианты являются десятичными дробями с k десятичными знаками после запятой, то, чтобы избежать действий с дробями, умножают первоначальные варианты на постоянное число $C=10^k$, т.е. переходят к *условным вариантам* $u_i = C \cdot x_i$. При этом выборочное среднее увеличится в C раз, а дисперсия - в C^2 раз, поэтому справедливы выражения:

$$u_i = C \cdot x_i \quad \rightarrow \quad \overline{x}_B = \frac{\overline{u}_B}{C}; \quad D_B(X) = \frac{D_B(u)}{C^2}. \tag{11}$$

Пример 2.1. Найти выборочное среднее, исправленную выборочную дисперсию и исправленное выборочное среднее квадратическое отклонение для выборок, заданных в примере 1.1.

x_i	3,2	4,4	5,0	6,7	7,5	8,1
n_i	3	5	3	5	1	3
w_i	0,15	0,25	0,15	0,25	0,05	0,15

Решение.

$$\overline{x}_B = \frac{3,2 \cdot 3 + 4,4 \cdot 5 + 5 \cdot 3 + 6,7 \cdot 5 + 7,5 \cdot 1 + 8,1 \cdot 3}{20} = 5,595;$$

$$s^2 = \frac{(3,2-5,595)^2 \cdot 3 + ... + (8-5,595)^2 \cdot 3}{19} = 2,84; \; s = \sqrt{2,84} = 1,69.$$

Замечание. В интервальном статистическом ряде вариантами следует считать середины частичных интервалов.

Другие характеристики вариационного ряда

Кроме выборочной средней и выборочной дисперсии применяются и другие характеристики вариационного ряда. Укажем главные из них.

Медианой *Me* называют варианту, которая делит вариационный ряд на две части, равные по числу вариант. Если число вариант нечетно, т. е. $n=2k+1$, то $Me = x_{k+1}$, если четно, т.е. $n = 2k$, то $Me = (x_k + x_{k+1})/2$.

Для примера 1.1: $n = 20$, $k=10$, $Me = (x_{10} + x_{11})/2 = (6,7 + 6,7)/2 = 6,7$.

Для интервального ряда медиане соответствует значение кумуляты частостей равное 0,5, поэтому используют формулу (12), которая получена из подобия треугольников, показанных на рис. 2.1

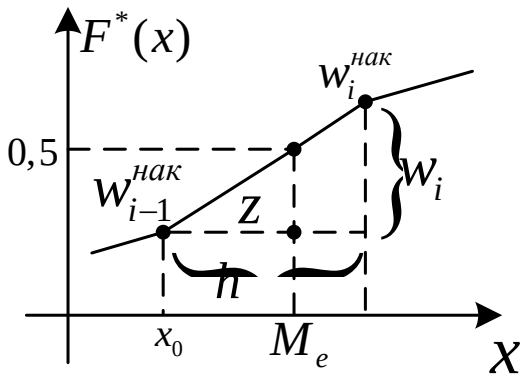

Рисунок 2.1 - К определению медианы *Me*

$$\frac{z}{h}=\frac{0,5-w_{i-1}^{нак}}{w_i};\qquad Me=x_0+z;$$

$$Me=x_0+h\frac{0,5-w_{i-1}^{нак}}{w_i},\qquad (12)$$

где x_0 - начало интервала, содержащего медиану; w_i, $w_{i-1}^{нак}$ - частость и накопленная частость медианного и предмедианного интервалов соответственно.

Для примера 1.2 (см. табл.1.2) получим по формуле (12):

$$Me=0,29+0,43\frac{0,5-0,478}{0,188}=0,34.$$

Модой *Mo* называют варианту, которая имеет наибольшую частоту.

Для интервального ряда используют формулу (13), которая получена из подобия треугольников, показанных на рис. 2.2.

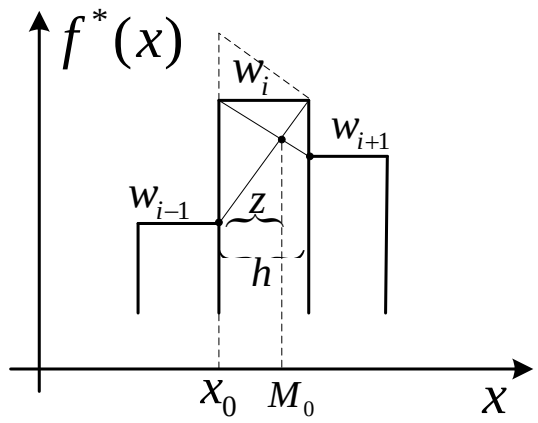

Рисунок 2.2 - К определению моды *Mo*

$$\frac{z}{h}=\frac{w_i-w_{i-1}}{(w_i-w_{i-1})+(w_i-w_{i+1})};$$

$$M_0=x_0+z;$$

$$M_0=x_0+h\frac{w_i-w_{i-1}}{2w_i-w_{i-1}-w_{i+1}},\qquad (13)$$

где x_0 - начало интервала, содержащего моду; w_{i-1}, w_i, w_{i+1} - частоты предмодального, модального и постмодального интервалов соответственно.

В примере 1.1 имеется две моды $M_{0\_1}=4,4$; $M_{0\_2}=6,7$.

Для примера 1.2 (см. табл.1.2) получим по формуле (13):

$$M_0=0,72+0,43\frac{0,203-0,188}{2\cdot 0,203-0,188-0,087}=0,769.$$

Размахом варьирования *R* называют разность между наибольшей и наименьшей вариантами:

$$R=x_{\max}-x_{\min}.$$

Коэффициент вариации *V* - безразмерная величина, которая характеризует в процентах долю выборочного СКО от выборочной средней:

$$V = \frac{\sigma_B}{\overline{x}} 100\%.$$

2.3. Вычисление точечных оценок в Excel

Для вычисления выборочного среднего (5) используется ***функция СРЗНАЧ***, обращение к которой имеет вид:

=СРЗНАЧ(*арг1; арг2; …; арг255*),

где *арг1; арг2; …; арг255* – числа или адреса ячеек (не более 225), содержащих числовые данные. Если ячейка содержит текстовые, логические значения или ячейка пуста, то такие ячейки игнорируются.

Для вычисления выборочной (6) и исправленной (8) дисперсий используются ***функции ДИСПР*** и ***ДИСП соответственно,*** обращение к которым имеет вид:

=ДИСПР(*арг1; арг2; …; арг255*);

=ДИСП(*арг1; арг2; …; арг255*).

Для вычисления суммы квадратов отклонений $\sum_{i=1}^{n}(x_i - \overline{x}_B)^2$ используется ***функция КВАДРОТКЛ***, обращение к которой имеет вид:

=КВАДРОТКЛ(*арг1; арг2; …; арг255*).

Для вычисления ***выборочного и исправленного СКО*** используются соответственно функции:

=СТАНДОТКЛОНП (*арг1; арг2; …; арг255*);

=СТАНДОТКЛОН (*арг1; арг2; …; арг255*).

Находят применение также следующие встроенные функции.

Функция ЭКСЦЕСС вычисляет оценку

$$\frac{n(n+1)}{(n-1)(n-2)(n-3)}\sum_{i=1}^{n}\left(\frac{x_i - \overline{x}_в}{d_в}\right)^2 - \frac{3(n-1)^2}{(n-2)(n-3)}$$

для характеристики *эксцесс* $\frac{\mu_4}{\sigma^4} - 3$, которая определяет островершинность или плосковершинность плотности распределения.

Функция МОДА вычисляет наиболее часто встречающееся значение в заданных аргументах функции, т.е. значение, встречающееся в выборке с максимальной частотой.

Если в заданных значениях аргументов ***нет повторяющихся значений***, то функция возвращает признак ошибки #Н/Д.

Функция МЕДИАНА вычисляет значение выборки, приходящееся на середину упорядоченной выборочной совокупности. Если выборка имеет четное число элементов, то значение функции будет равно среднему двух значений, находящихся по середине упорядоченной выборочной совокупности.

Функция СКОС вычисляет оценку

$$\frac{n}{(n-1)(n-2)}\sum_{i=1}^{n}\frac{\left(x_i-\overline{x}_в\right)^3}{d_в^{3/2}}$$

для характеристики асимметрии $\frac{\mu_3}{\sigma^3}$, которая для симметричной плотности распределения равна 0.

Основные характеристики положения, разброса и асимметрии можно также вычислить, используя вкладку ***Данные - Анализ данных - Описательная статистика*** *пакета анализа*.

В появившемся диалоговом окне Описательная статистика параметр *Уровень надежности* включается, если необходимо вычислить доверительный интервал для математического ожидания с задаваемым ($в\,\%$) уровнем надежности γ. *Уровень надежности* – определяет величину $\Delta_{\overline{x}}$, от которой зависит доверительный интервал для математического ожидания, имеющий вид

$$\left[\overline{x}_в-\Delta_{\overline{x}},\overline{x}_в+\Delta_{\overline{x}}\right],$$

где $\overline{x}_в$ – выборочное среднее (подробнее см. Интервальные оценки).

Назначение остальных параметров достаточно очевидно.

2.4. Интервальные оценки

Точечная оценка при малом объеме выборки может существенно отличаться от оцениваемого параметра, поэтому важно знать, насколько истинное значение параметра может отклоняться от найденной точечной оценки. Интервал вида $|\theta-\theta^*|<\delta$, где Θ - истинное значение оцениваемого параметра, а Θ^* - его точечная оценка, называется *доверительным интервалом*, а вероятность $\gamma=p\left(|\theta-\theta^*|<\delta\right)$ - *доверительной вероятностью* или *надежностью*.

Границы доверительного интервала являются случайными величинами и с вероятностью γ накрывают истинное значение оцениваемого параметра.

Для построения доверительного интервала требуется знать закон распределения исследуемой случайной величины $f\left(\theta^*\right)$, как показано на рисунке 2.3.

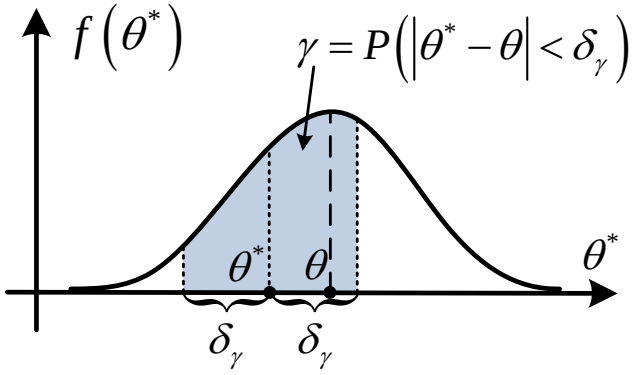

Рисунок 2.3 - К определению интервальной оценки θ^*

Интервальные оценки математического ожидания нормального распределения

Пусть генеральная совокупность X распределена ***по нормальному закону*** $N(a,\sigma)$, причем ***параметр*** σ ***известен***, а параметр $a = M(X)$ требуется оценить с надежностью γ. По теореме о распределении выборочных характеристик случайная величина $Z = \dfrac{\left(\bar{X}_в - a\right)\sqrt{n}}{\sigma}$ распределена по закону $N(0,1)$ с плотностью вероятностей $f\left(z\right) = \dfrac{1}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}$ и называется *z-статистикой*.

На рисунке 2.8 изображены графики плотностей случайной величины $\bar{X}_в$, распределенной по $N\left(a, \dfrac{\sigma}{\sqrt{n}}\right)$ и *z-статистики,* распределенной по $N\left(0,1\right)$.

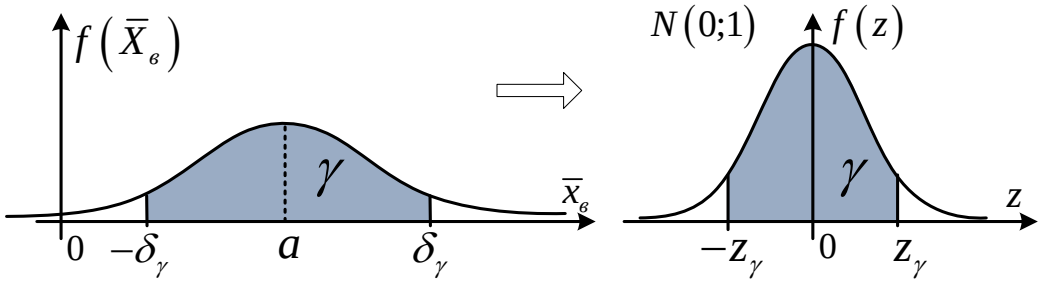

Рисунок 2.8 - К построению доверительных интервалов

Зададимся требуемой надежностью γ оценки a и, полагая известными σ и объем выборки n, определим точность оценки δ_γ - ширину доверительного интервала

$$\gamma = P\left(\left|\overline{X}_в - a\right| < \delta_\gamma\right) = P\left(\left|\frac{\left(\overline{X}_в - a\right)\sqrt{n}}{\sigma}\right| < \delta_\gamma \frac{\sqrt{n}}{\sigma}\right) = P\left(|Z| < z_\gamma\right), \qquad (14)$$

где $\delta_\gamma = z_\gamma \frac{\sigma}{\sqrt{n}}$ - искомый доверительный интервал.

Выражение (14) эквивалентно следующему

$$P\left(\overline{X}_в - z_\gamma \frac{\sigma}{\sqrt{n}} < a < \overline{X}_в + z_\gamma \frac{\sigma}{\sqrt{n}}\right) = \gamma, \qquad (15)$$

которое определяет ***доверительный интервал (интервальную оценку) для математического ожидания a*** с точностью $\delta_\gamma = z_\gamma \frac{\sigma}{\sqrt{n}}$ и надежностью γ.

Значение $z = z_\gamma$ находится с использованием интегральной функции Лапласа $\Phi(z)$, представленной в Приложении 2. Действительно,

$$P\left(-z < N(0,1) < z\right) = \Phi(z) - \Phi(-z) = 2\Phi(z) = \gamma. \qquad (16)$$

По табл. П2 определяем значение z, удовлетворяющее уравнению

$$\Phi(z) = \frac{\gamma}{2}; \quad z = \Phi^{-1}\left(\frac{\gamma}{2}\right). \qquad (17)$$

Пример 2.1. Дана выборка значений нормально распределенной случайной величины: 2, 3, 3, 4, 2, 5, 5, 5, 6, 3, 6, 3, 4, 4, 4, 6, 5, 7, 3, 5. Найти с доверительной вероятностью γ = 0,95 границы доверительных интервалов для математического ожидания, если известно СКО распределения $\sigma = 1{,}37$.

Решение.

Поскольку распределение нормальное и дисперсия известна $\sigma^2 = 1{,}37^2 = 1{,}88$, то для оценки $\bar{x}_B$ используем z-распределение. Найдем n=20, $\bar{x}_B$ = 4,25. По таблице Прил.2 определим аргумент z=1,96, при котором функции Лапласа $\Phi(z) = \frac{0{,}95}{2} = 0{,}475$. Тогда из формулы (15)

$$4{,}25 - 1{,}96\frac{1{,}37}{\sqrt{20}} < a < 4{,}25 + 1{,}96\frac{1{,}37}{\sqrt{20}};$$

$$4{,}25 - 0{,}6 < a < 4{,}25 + 0{,}6; \quad 3{,}65 < a < 4{,}85.$$

Минимальный объем выборки n, при котором оценку математического ожидания *a* можно получить с заданной надежностью γ и точностью δ_γ.

Интервальная оценка *a* зависит от трех взаимосвязанных параметров: надежности γ, точности δ_γ и объема выборки *n*. Задаваясь двумя из них можно определить оставшийся.

Пусть $\delta_\gamma = \left|\overline{X}_в - a\right|$. Тогда на основании формулы (14) ***минимальный объем выборки n***, гарантирующий оценку математического ожидания *a* с заданной надежностью γ и точностью δ_γ определяется неравенствами

$$\delta_\gamma \geq z_\gamma \frac{\sigma}{\sqrt{n}}, \qquad \delta_\gamma \sqrt{n} \geq z_\gamma \sigma, \qquad n \geq \left(\sigma \frac{z_\gamma}{\delta_\gamma} \right)^2. \tag{18}$$

Пример 2.2. Найти минимальный объем выборки, при котором с надежностью 0,95 точность оценки математического ожидания *a* генеральной совокупности по выборочной средней равна $\delta = 0{,}5$, если известно СКО $\sigma = 1{,}37$ нормально распределенной генеральной совокупности.

Решение.

По таблице Прил.2 определим аргумент $z = 1{,}96$, при котором функции Лапласа $\Phi(z) = \frac{0{,}95}{2} = 0{,}475$. Тогда по формуле (18)

$$n \geq \left(\sigma \frac{z_\gamma}{\delta_\gamma} \right)^2 = \left(1{,}37 \frac{1{,}96}{0{,}5} \right)^2 = \left(5{,}37\right)^2 = 28{,}84. \quad \text{Т.о., } n_{\min} = 29.$$

При неизвестном СКО и объеме выборки $n < 30$, для нормального ЗР, на основании теоремы **Ошибка! Источник ссылки не найден.** имеем

$$\frac{\left(\overline{X}_в - a\right)\sqrt{n}}{s} = T_{n-1}.$$

Поэтому доверительный интервал для математического ожидания при заданной надежности γ определяется так:

$$\overline{x}_B - \frac{t_\gamma \cdot s}{\sqrt{n}} < a < \overline{x}_B + \frac{t_\gamma \cdot s}{\sqrt{n}}. \tag{19}$$

Здесь *s* – исправленное выборочное среднее квадратическое отклонение, а $t_\gamma = t_\gamma(\gamma, k)$ – критическая точка распределения Стьюдента, определяемая

выражением $P\left(|T_k| < t(\gamma,k)\right) = \gamma$, где $k = n-1$ – количество степеней свободы. Значения t_γ можно найти из таблицы Приложения 3 по известным n и γ.

При неизвестном СКО и объеме выборки $n \geq 30$, для нормального ЗР, полагают, что распределение Стьюдента несущественно отличается от нормального. Поэтому для оценки математического ожидания применяют формулу (15), где вместо σ используют s.

Пример 2.3. Дана выборка значений нормально распределенной случайной величины: 2, 3, 3, 4, 2, 5, 5, 5, 6, 3, 6, 3, 4, 4, 4, 6, 5, 7, 3, 5. Найти с доверительной вероятностью γ = 0,95 границы доверительного интервала для математического ожидания.

Решение.

Поскольку объем выборки небольшой n = 20 и дисперсия не известна, то для оценки $\overline{x}_B$ используем распределение Стьюдента. Найдем $\overline{x}_B$ = 4,25, s = 1,37. По таблице Прил.3 определим $t_\gamma(0{,}95;\ 19) = 2{,}093$. Тогда по формуле (19)

$$4{,}25 - \frac{2{,}093 \cdot 1{,}37}{\sqrt{20}} < a < 4{,}25 + \frac{2{,}093 \cdot 1{,}37}{\sqrt{20}}; \quad 3{,}64 < a < 4{,}86$$

- доверительный интервал для математического ожидания.

Интервальные оценки дисперсии нормального распределения

Как и при построении интервальных оценок для математического ожидания, в данном случае также необходимо определить статистику (функцию от наблюдаемых вариант), распределение которой было бы известно и включало бы оцениваемый параметр σ.

В соответствии с теоремой о распределении выборочных характеристик **Ошибка! Источник ссылки не найден.** такой статистикой может быть случайная величина $\frac{nD_в}{\sigma^2}$ или $\frac{(n-1)S^2}{\sigma^2}$, распределенная по закону χ^2_{n-1} с $(n-1)$ степенями свободы. Заметим, что распределение χ^2, в отличие от распределения Стьюдента, не является симметричным, поэтому для доверительного интервала целесообразно выбрать два предела $\chi^2_{лев,\gamma}$ и $\chi^2_{пр,\gamma}$ так, чтобы площади "хвостов" были равными $\frac{\alpha}{2}$ [5]

$$P\left(\chi^2_{n-1} < \chi^2_{лев,\gamma}\right) = P\left(\chi^2_{n-1} > \chi^2_{пр,\gamma}\right) = \frac{\alpha}{2}, \qquad (20)$$

где $\alpha = 1-\gamma$, α, γ – уровень значимости и надежность оценки;

$\chi^2_{лев,\gamma}$ – критическая точка χ^2_{n-1}-распределения уровня $1-\alpha/2$ (или квантиль уровня $\alpha/2$), $\chi^2_{пр,\gamma}$ – критическая точка уровня $\alpha/2$ (или квантиль уровня $1-\alpha/2$). Таблица критических точек представлена в Приложении 4.

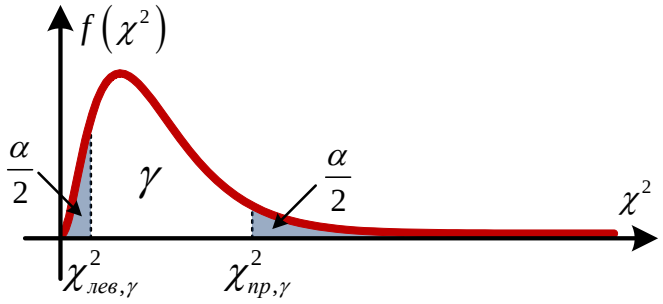

Рисунок 2.9 - К построению доверительных интервалов

На основании равенств

$$P\left(\chi^2_{лев,\gamma} < \frac{nD_в}{\sigma^2} < \chi^2_{пр,\gamma}\right) = \gamma \quad \text{или} \quad P\left(\frac{nD_в}{\chi^2_{лев,\gamma}} > \sigma^2 > \frac{nD_в}{\chi^2_{пр,\gamma}}\right) = \gamma, \qquad (21)$$

интервал

$$\frac{nD_в}{\chi^2_{пр,\gamma}} < \sigma^2 < \frac{nD_в}{\chi^2_{лев,\gamma}} \quad \text{или} \quad \frac{n-1}{\chi^2_{пр,\gamma}} s^2 < \sigma^2 < \frac{n-1}{\chi^2_{лев,\gamma}} s^2 \qquad (22)$$

является интервальной оценкой для σ^2 ***с надежностью*** γ**.**

Границы интервалов (22) являются случайными величинами и с вероятностью γ накрывают неизвестную дисперсию σ^2.

Пример 2.4. По выборке объема *n* = 20 из нормально распределенной генеральной совокупности вычислено значение выборочной дисперсии $D_в = 1{,}5$. Построить интервальную оценку для параметра σ^2 с надежностью γ = 0,96.

Решение. Значения $\chi^2_{лев,\gamma}$, $\chi^2_{пр,\gamma}$ находим из условий (20):

$$\alpha = 0{,}04; \quad P\left(\chi^2_{19} > \chi^2_{лев,\gamma}\right) = 0{,}98; \quad P\left(\chi^2_{19} > \chi^2_{пр,\gamma}\right) = 0{,}02.$$

Т.е. $\chi^2_{лев,\gamma}$ есть критическая точка χ^2-распределения с 19 степенями свободы уровня 0,98, а $\chi^2_{пр,\gamma}$ – критическая точка уровня 0,02.

По табл. Приложения 4 критических точек χ^2-распределения находим

$$\chi^2_{лев,\gamma} = 8{,}6; \qquad \chi^2_{пр,\gamma} = 33{,}7.$$

Тогда интервальная оценка σ^2 принимает вид (22)

$$\left(\frac{20}{33,7}D_в;\ \frac{20}{8,6}D_в\right)=\left(0,59D_в;\ 2,33D_в\right).$$

Подставляя вычисленное значение $D_в=1,5$, получаем

$$0,89<\sigma^2<3,488,$$

откуда оценка СКО $0,94<\sigma<1,868.$

Интервальная оценка вероятности события

Точечной оценкой вероятности p события является частность $w=m/n$, где n – общее число независимых испытаний, а m – число испытаний, в которых произошло событие A.

Зададимся надежностью интервальной оценки γ и найдем числа $p_{лев,\gamma}$, $p_{пр,\gamma}$ такие, чтобы выполнялось соотношение

$$P\left(p_{лев,\gamma}<p<p_{пр,\gamma}\right)=\gamma. \qquad (23)$$

Интервальная оценка вероятности при большом числе испытаний. Если $n>30$ и $np>10$, то распределение случайной величины $w=\frac{m}{n}$ можно аппроксимировать нормальным распределением $N\left(p,\sqrt{pq/n}\right)$. Следовательно, при этих же условиях распределение величины $\frac{(w-p)}{\sqrt{pq/n}}$ близко к нормальному с нулевым математическим ожиданием и единичной дисперсией, т.е.

$$\frac{w-p}{\sqrt{pq/n}}=N(0,1).$$

По аналогии с (14), найдем такое число z_γ, для которого справедливо равенство

$$P\left(-z_\gamma<\frac{w-p}{\sqrt{pq/n}}<z_\gamma\right)=\gamma. \qquad (24)$$

Это число z_γ является корнем уравнения $\Phi(z_\gamma)=\gamma/2$, где $\Phi(z)$ – функция Лапласа, и корень может быть найден с помощью табл. П2.

Неравенство, стоящее в скобках выражения (24), разрешим относительно p. Для этого неравенство перепишем в виде эквивалентного неравенства

$\left|\frac{w-p}{\sqrt{pq/n}}\right| < z$. Возведем в квадрат, в результате получим $(w-p)^2 < \frac{p(1-p)}{n} z^2$.

Далее, возведя в квадрат $(w-p)$ и перенеся все члены влево, получим

$$\left(1+\frac{z^2}{n}\right)p^2 - \left(2w+\frac{z^2}{n}\right)p + w^2 < 0.$$

Корни p_1 и p_2 этого квадратного трехчлена ***являются границами интервальной оценки*** (23) ***вероятности события*** и определяются выражениями

$$p_1 = p_{лев,\gamma} = \frac{n}{z^2+n}\left[w + \frac{z^2}{2n} - z\sqrt{\frac{w(1-w)}{n} + \left(\frac{z}{2n}\right)^2}\right],$$
$$p_2 = p_{пр,\gamma} = \frac{n}{z^2+n}\left[w + \frac{z^2}{2n} + z\sqrt{\frac{w(1-w)}{n} + \left(\frac{z}{2n}\right)^2}\right], \quad (25)$$

где n — общее число испытаний; m — число появлений события; w — относительная частота; z — значение аргумента функции Лапласа (приложение 2), при котором $\Phi(z) = \frac{\gamma}{2}$ (γ — заданная надежность).

Если $n >> 100$, то в формулах (25) слагаемым $\frac{z}{2n}$ можно пренебречь, тогда для вычисления p_1, p_2 можно использовать приближенные формулы, аналогичные (15):

$$p_1 = w - z\sqrt{\frac{w(1-w)}{n}},\ p_2 = w + z\sqrt{\frac{w(1-w)}{n}}. \quad (26)$$

Пример 2.5. Событие A в серии из n = 100 испытаний произошло m = 78 раз. Построить интервальную оценку для вероятности p события с надежностью $\gamma = 0,9$.

Решение. Значение точечной оценки вероятности p равно $w = 78/100 = 0,78$. По табл. П2 определяем для $\Phi(z) = 0,9/2 = 0,45$ $z = 1,64$ и вычисляем по формулам (25) значения p_1, p_2 при $w = 0,78$: $p_1 = 0,705$, $p_2 = 0,848$. Таким образом, доверительный интервал для вероятности p события A следующий: (0,705; 0,848).

Интервальная оценка вероятности при малом числе испытаний (n<30). При малом числе испытаний n предположение о приближенном распределении случайной величины m по нормальному закону становится несправедливым. Для описания распределения величины m необходимо использовать формулу Бернулли:

$$P(m=x)=C_n^x p^x (1-p)^{n-x}, \quad x=0,1,...,n.$$

Можно показать, что граничные точки интервальной оценки (23) являются решениями следующих нелинейных уравнений:

$$\sum_{x=0}^{m-1} C_n^x p_{лев,\gamma}^x (1-p_{лев,\gamma})^{n-x} = \frac{1+\gamma}{2}; \qquad \sum_{x=0}^{m} C_n^x p_{пр,\gamma}^x (1-p_{пр,\gamma})^{n-x} = \frac{1-\gamma}{2},$$

где γ – надежность интервальной оценки.

Корни этих уравнений могут быть найдены одним из численных методов решения нелинейных уравнений. Кроме этого, существуют специальные таблицы для нахождения $p_{лев,\gamma}, p_{пр,\gamma}$. В данном пособии они не приводятся.

2.5. Вычисление границ доверительных интервалов в Excel

Вычисление величины z_γ, входящей в доверительный интервал $P\left(|Z|<z_\gamma\right)$ оценки математического ожидания (14).

В Excel реализована не функция Лапласа (см. Прил.2), а интегральная функции стандартного нормального распределения $F(z)=P\left(Z<z\right)$, где $\left(-5<z<5\right)$. Поэтому величина z_γ (14) вычисляется с помощью функции НОРМСТОБР(p) следующим образом:

$z_\gamma = \text{НОРМСТОБР}\left(1-\alpha/2\right)$, где γ – надежность интервальной оценки.

Вычисление величины $\delta_\gamma = z_\gamma \frac{\sigma}{\sqrt{n}}$ (15) осуществляется с помощью функции ДОВЕРИТ:

$$\delta_\gamma = z_\gamma \frac{\sigma}{\sqrt{n}} = \text{ДОВЕРИТ}(\alpha;\sigma;n),$$

где $\alpha=1-\gamma$, σ – известное СКО, n – объем выборки.

Вычисление $t_\gamma(\gamma,k)$ критической точки распределения Стьюдента (19), определяемой выражением $P\left(|T_k|<t\left(\gamma,k\right)\right)=\gamma,$ осуществляют с использованием функции СТЬЮДРАСПОБР, обращение к которой имеет вид:

$$t_{\gamma}(\gamma, k) = \text{СТЬЮДРАСПОБР}(\alpha; k),$$

где $\alpha = 1 - \gamma$, $k = n - 1$ – *число степеней свободы*.

Вычисление критических точек распределения Пирсона $\chi^2_{лев,\gamma}$, $\chi^2_{пр,\gamma}$, входящих в доверительный интервал (22), для дисперсии σ^2 выполняется с использованием функции ХИ2ОБР:

$$\chi^2_{лев,\gamma} = \text{ХИ2ОБР}\left(1 - \alpha/2; k\right); \qquad \chi^2_{пр,\gamma} = \text{ХИ2ОБР}\left(\alpha/2; k\right),$$

где $\alpha = 1 - \gamma$, γ – надежность интервальной оценки.

Типовой вариант задания

1. Первичная обработка информации. Оценки параметров.

1.1. i-му варианту соответствуют только элементы выборки "x" соответствующей строки (объем выборки при этом $n=16$).

1.2. Построить интервальный вариационный ряд; гистограмму относительных частот; эмпирическую функцию распределения.

1.3. Определить выборочное среднее, выборочную и исправленную дисперсии, СКО, моду, медиану, вариационный размах и коэффициент вариации.

1.4. Полагая закон распределения вариант нормальным, найти интервальную оценку математического ожидания с надежностью γ_m (см. таблицу).

1.5. Найти минимальный объем выборки, обеспечивающий с надежностью γ_m точность δ_m оценки математического ожидания.

1.6. Найти интервальные оценки дисперсии и СКО генеральной совокупности с надежностью γ_D (см. таблицу).

Вариант	Номер измерения															
	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
x	2,3	2,1	4,6	4,9	3,7	5	2,8	1,6	4,8	3,7	1,1	2,4	4,3	4,5	3,2	2,4

Интервал. оц.			Проверка гипотез				
γ_m	δ_m	γ_D	α_m	зн	a_0	D_0	α_S
0,95	*0,16*	*0,95*	*0,05*	*<4,8*		*0,75*	

ЛИТЕРАТУРА

1. **Вентцель Е.**С. Теория вероятностей: учебник для студентов вузов. – 10 изд. – М.: Издательский центр «Академия», 2005. – 576с.

2. **Гмурман В.Е.** Теория вероятностей и математическая статистика: уч. пособие. – 12-е изд. – М.: Высшее образование, 2008. - 479с.

3. **Кремер Н.Ш.** Теория вероятностей и математическая статистика: учебник для вузов. – М.: ЮНИТИ-ДАНА, 2000. - 543с.

4. **Гмурман В.Е.** Руководство к решению задач по теории вероятностей и математической статистике: уч. пособие для студентов вузов. – 4-е изд. – М.: Высшая школа, 1997. - 400с.

5. **Письменный Д.Т.** Конспект лекций по теории вероятностей, математической статистике и случайным процессам. – 3-е изд. – М.: Айрис-пресс, 2008. – 288с.

6. **Гулай Т.А., Долгополова А.Ф., Литвин Д.Б., Мелешко С.В.** Теория вероятностей и математическая статистика. - 2-е изд. - Ставрополь : Агрус, 2013. - 256с.

7. **Литвин Д.Б., Мелешко С.В.** Элементы теории вероятностей: Учебное пособие. – Ставрополь: Сервисшкола, 2017. – 86 с.