МИНОБРНАУКИ РОССИИ

Федеральное государственное бюджетное образовательное учреждение высшего образования
«Юго-Западный государственный университет»
(ЮЗГУ)

Кафедра высшей математики

ЭЛЕМЕНТЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Методические указания по выполнению модуля «Элементы математической статистики и корреляционного анализа»

Курск 2018

УДК 519.22

Составители: О.А. Бредихина, С.В. Шеставина

Рецензент

Кандидат технических наук, доцент кафедры высшей математики

Н.А. Моргунова

Элементы математической статистики: методические указания по выполнению модуля «Элементы математической статистики и корреляционного анализа» / Юго-Зап. гос. ун-т; сост.: О.А. Бредихина, С.В. Шеставина. – Курск, 2018. – 28 с.

Методические указания содержат теоретические сведения по теме «Элементы математической статистики», подробно разобраны типовые задачи.

Методические указания предназначены для студентов всех направлений и специальностей, изучающих математическую статистику.

Текст печатается в авторской редакции

Подписано в печать 1.02.18. _. Формат 60х84 1/16.

Усл. печ. л. 1,4. Уч.-изд. л. 1,3. Тираж 100 экз. Заказ 361 . Бесплатно.

Юго-западный государственный университет.

305040 Курск, ул. 50 лет Октября, 94.

Содержание

Введение

В образовательном процессе математическая статистика традиционно считается наиболее сложным для восприятия предметом. Предлагаемая работа ставит своей целью помочь тем, кто осваивает этот раздел математики.

В ней содержатся основные положения математической статистики, а также разобраны задачи с подробным решением по указанной тематике.

Данное пособие является приложением к модулю 20 «Элементы математической статистики и корреляционного анализа».

Авторы надеются, что это методическое издание поможет студентам в самостоятельной работе по выполнению модуля и изучению данного материала.

1. Статистическое распределение выборки. Графическое изображение вариационных рядов

Пусть из генеральной совокупности извлечена выборка *объёма* n, в которой значения x_1 некоторого исследуемого признака X_1 наблюдалось n_1 раз, значения $x_2 - n_2$ раз, значения $x_m - n_m$ раз. Значения x_i называются *вариантами*, а n_i – их *частотами.*

Вариационным рядом называется ранжированный в порядке возрастания (или убывания) ряд вариант с соответствующими им частотами.

Отношения частот вариант к объёму выборки

$$\omega_i = \frac{n_i}{n}$$

называются *относительными частотами.* При этом $\sum_{i=1}^{m} n_i = n$.

Перечень вариант и соответствующих им частот (относительных частот) называется *статистическим распределением выборки* или *статистическим рядом.* Здесь имеется аналогия с законом распределения случайной величины: в теории вероятностей – это соответствие между возможными значениями случайной величины и их вероятностями, а в математической статистике – это соответствие между наблюдаемыми вариантами и их частотами (относительными частотами). Заметим, что сумма относительных частот равна единице, т. е. $\sum_{i=1}^{k} \omega_i = 1$.

Вариационный ряд называется *дискретным*, если любые его варианты отличаются на постоянную величину, и *непрерывным (интервальным)*, если варианты могут отличаться один от другого на сколь угодно малую величину.

Для *преобразования дискретного статистического ряда в интервальный* необходимо установить величину интервалов, получить шкалу интервалов и в соответствии с этой шкалой сгруппировать данные.

Для определения величины интервала используется *формула Стерджесса:*

$$h = \frac{x_{\max} - x_{\min}}{1 + 3{,}322 \cdot \lg n},$$

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

$1 + 3{,}322 \cdot \lg n = 1 + \log_2 n$ – число интервалов.

Шкала интервалов формируется следующим образом:

$a_1 = x_{\min}$,

$a_2 = a_1 + h$,

$a_3 = a_2 + h$,

…

Границы интервалов формируются до тех пор, пока не превысят $x_{\max}$: $[a_1; a_2)$, $[a_2; a_3)$, …, $[a_{m-1}; a_m]$, где $a_m \geq x_{\max}$.

Во второй строке статистического ряда записывается количество наблюдений, попавших в каждый интервал.

Для изображения вариационных рядов обычно используются полигон, гистограмма, кумулятивная кривая.

1. Дискретный вариационный ряд графически изображается полигоном.

Каждую пару значений $(x_i; n_i)$ из распределения выборки можно трактовать как точку на координатной плоскости. Точно так же можно рассматривать и пары значений $(x_i; \omega_i)$ относительного распределения выборки.

Полигоном частот называется ломаная, отрезки которой соединяют точки $(x_i; n_i)$.

Полигоном относительных частот называется ломаная, отрезки которой соединяют точки $(x_i; \omega_i)$.

2. Интервальный вариационный ряд графически изображается гистограммой.

Гистограммой частот называется ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длины h, а высоты равны отношению $\frac{n_i}{h}$ (*плотность частоты*).

Геометрический смысл гистограммы частот: площадь гистограммы частот равна сумме всех частот, т.е. объёму выборки n.

Гистограммой относительных частот называется ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длины h, а высоты равны отношению $\frac{\omega_i}{h}$ (*плотность относительной частоты*).

Геометрический смысл гистограммы относительных частот: площадь гистограммы относительных частот равна сумме всех частот, т.е. единице.

3. *Кумулятивная кривая (кумулята, кривая накопленных частот (относительных частот))* строится как для дискретных, так и для интервальных вариационных рядов.

Для *дискретных* вариационных рядов кумулятивная кривая – ломаная, соединяющая точки $\left(x_i; n_i^{нак}\right)$ или $\left(x_i; \omega_i^{нак}\right)$, $\omega_i^{нак} = \frac{n_i^{нак}}{n}$, где $n_i^{нак}$ – накопленная частота.

Для *интервального* вариационного ряда по оси абсцисс откладываются интервалы. Верхним границам интервалов соответствуют накопленные частоты, а нижней границе первого интервала – накопленная частота, равная нулю.

Пример решения задания 1

Имеются данные о стаже рабочих цеха: 6, 6, 10, 10, 7, 2, 2, 5, 8, 8, 12, 9, 10, 10, 7, 7, 6, 7, 2, 3. Построить дискретный и интервальный вариационные ряды и изобразить их графически: построить полигон, гистограмму, кумулятивную кривую.

Решение

1) Построение дискретного вариационного ряда.

Находится объём выборки, то есть n – общее количество чисел (у нас $n = 20$).

Составляется таблица, где x_i – варианты (числа из условия), расположенные в порядке возрастания, n_i – сколько раз встречается каждое число.

Дискретный вариационный ряд

x_i	2	3	5	6	7	8	9	10	12
n_i	3	1	1	3	4	2	1	4	1

Контроль: $\sum_{i=1}^{9} n_i = n$, то есть 3+1+1+3+4+2+1+4+1=20 – верно.

2) Построение интервального вариационного ряда

Находится величина интервала:

$$h = \frac{x_{\max} - x_{\min}}{1 + 3{,}322 \cdot \lg n} = \frac{12 - 2}{1 + 3{,}322 \cdot \lg 20} = \frac{10}{5{,}3} = 1{,}9.$$

Округление величины h производится следующим образом: если в условии даны числа, большие 100, то округляем до целых, если числа до 100, то оставляем один знак после запятой, а если даны очень маленькие числа, то округляем до сотых.

Далее шкала интервалов формируется так:

$a_1 = x_{i\min} = 2$,

$a_2 = a_1 + h = 2 + 1{,}9 = 3{,}9$,

$a_3 = a_2 + h = 3{,}9 + 1{,}9 = 5{,}8$,

$a_4 = 7{,}7$,

$a_5 = 9{,}6$,

$a_6 = 11{,}5$,

$a_7 = 13{,}4 > x_{\max} = 12$.

Составляется таблица, в которой первая строка имеет вид: $[a_1; a_2)$, $[a_2; a_3)$, … $[a_6; a_7]$. Вторая строка формируется так: $n_{i_{нов}}$ – общая сумма частоты встреч всех чисел дискретного ряда, попадающих в соответствующий интервал.

Интервальный вариационный ряд

Интервал	[2;3,9)	[3,9;5,8)	[5,8;7,7)	[7,7;9,6)	[9,6;11,5)	[11,5;13,4]
$n_{i_{нов}}$	3+1=4	1	3+4=7	2+1=3	4	1

3) Построение полигона частот по дискретному вариационному ряду.

Дискретный вариационный ряд имеет вид:

x_i	2	3	5	6	7	8	9	10	12
n_i	3	1	1	3	4	2	1	4	1

Соединив точки с координатами $(x_i; n_i)$, получим искомый график.

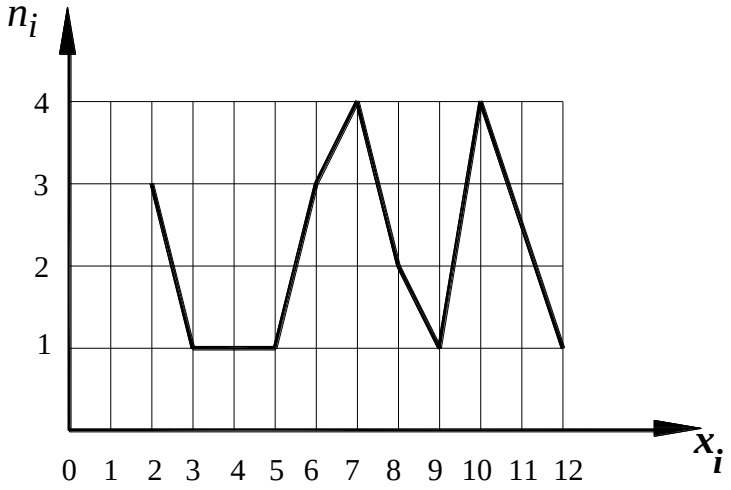

Рис. 1 Полигон частот

4) Построение гистограммы частот по интервальному вариационному ряду.

Дополним интервальный вариационный ряд. Последняя строка таблицы необходима для дальнейшего построения гистограммы частот.

Интервал	[2;3,9)	[3,9;5,8)	[5,8;7,7)	[7,7;9,6)	[9,6;11,5)	[11,5;13,4]
$n_{i\,нов}$	4	1	7	3	4	1
$\frac{n_{i\,нов}}{h}$	$\frac{4}{1,9}=2,1$	$\frac{1}{1,9}=0,5$	$\frac{7}{1,9}=3,7$	$\frac{3}{1,9}=1,6$	$\frac{4}{1,9}=2,1$	$\frac{1}{1,9}=0,5$

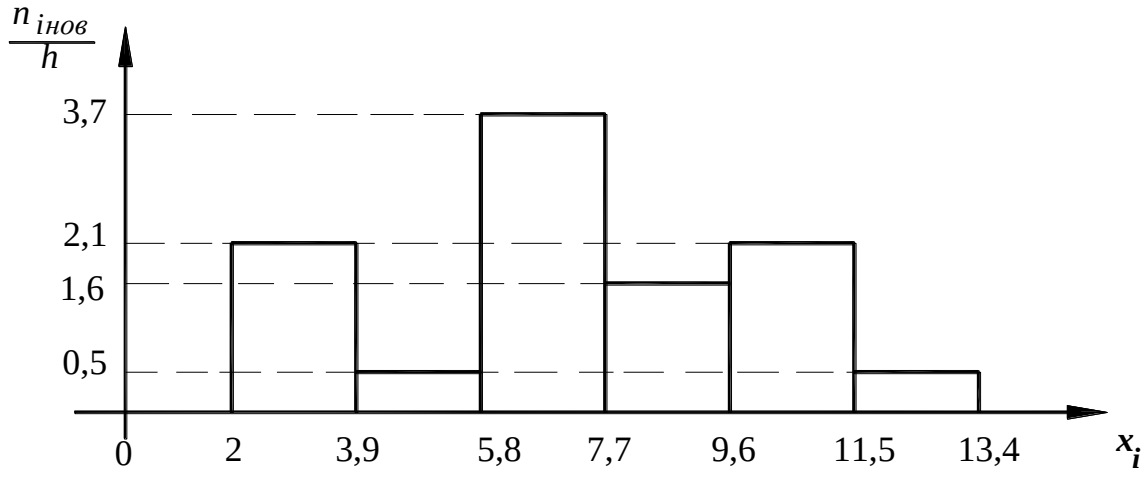

Рис. 2 Гистограмма частот

5) Построение кумулятивной кривой для дискретного вариационного ряда.

Дополним дискретный вариационный ряд. Последняя строка таблицы необходима для дальнейшего построения кумулятивной кривой.

x_i	2	3	5	6	7	8	9	10	12
n_i	3	1	1	3	4	2	1	4	1
$\frac{n_i^{нак}}{n}$	0,15	0,20	0,25	0,40	0,60	0,70	0,75	0,95	1

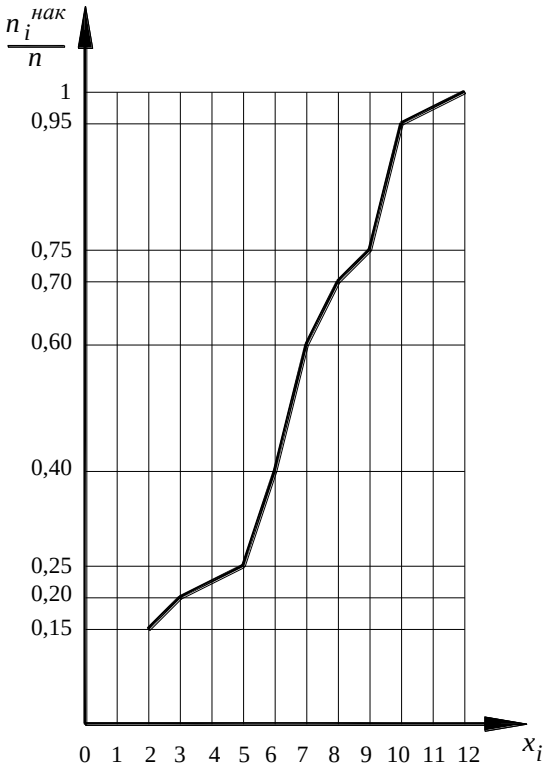

Рис. 3 Кумулятивная кривая для дискретного вариационного ряда

6) Построение кумулятивной кривой для интервального вариационного ряда.

Интервал	[2;3,9)	[3,9;5,8)	[5,8;7,7)	[7,7;9,6)	[9,6;11,5)	[11,5;13,4]
$n_{i_{нов}}$	4	1	7	3	4	1
$\frac{n_{i\,нов}^{нак}}{n}$	0,20	0,25	0,60	0,75	0,95	1

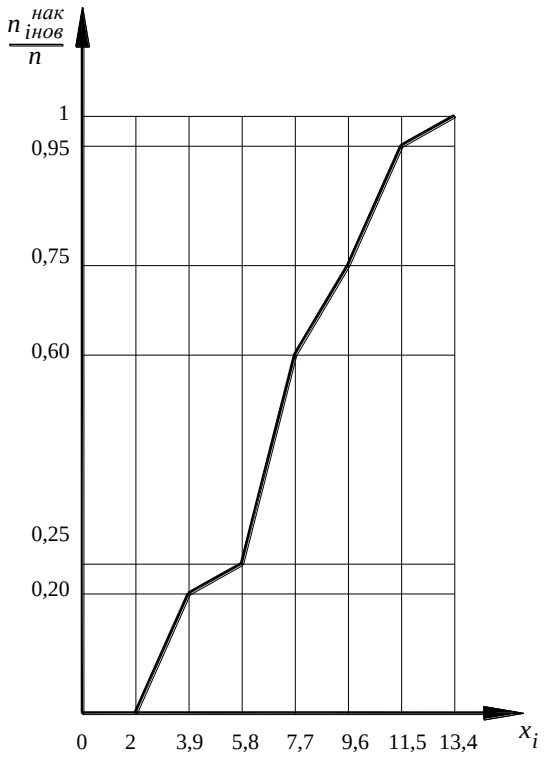

Рис. 4 Кумулятивная кривая для интервального вариационного ряда

2. Точечные оценки параметров распределения

Числовые характеристики всей генеральной совокупности называются *параметрами*. Так как всю генеральную совокупность изучить достаточно часто не представляется возможным, о параметрах судят по выборочным характеристикам.

На основании выборочных данных можно получить лишь приближенное значение параметра, которое является его *оценкой.*

Обозначим $\tilde{\Theta}_n$ – точечная оценка для параметра Θ генеральной совокупности.

Многократные выборки одинакового объёма дадут различные значения параметра Θ. В этом случае возникает проблема выбора наилучшей оценки. Чтобы выбранная оценка была наилучшей, она должна удовлетворять свойствам несмещённости, эффективности и состоятельности.

1. Оценка $\tilde{\Theta}_n$ называется *несмещённой*, если её математическое ожидание равно оцениваемому параметру, т.е. $M\left[\tilde{\Theta}_n\right]=\Theta$.

2. $\tilde{\Theta}_n$ называется *эффективной*, если она имеет наименьшую дисперсию при заданном объёме выборки.

3. *Состоятельной* называется оценка, которая при $n \to \infty$ стремится по вероятности к оцениваемому параметру $\lim\limits_{n\to\infty} P\left(\left|\tilde{\Theta}_n - \Theta\right| < \varepsilon\right) = 1$.

Репрезентативная выборка – это такая выборка, в которой все основные признаки генеральной совокупности, из которой извлечена данная выборка, представлены приблизительно в той же пропорции или с той же частотой, с которой данный признак выступает в этой генеральной совокупности. При достаточно большом значении объёма выборки нет разницы в выборе смещённой или несмещённой оценки параметра, при малом n необходимо рассматривать несмещённую оценку, поскольку выборка будет репрезентативной, то есть полностью представлять генеральную совокупность.

Для выборки можно определить ряд числовых характеристик: выборочное среднее, выборочная дисперсия, выборочное среднеквадратическое отклонение, размах выборки, асимметрия, эксцесс.

Пусть статистическое распределение выборки объёма *n* имеет вид:

x_i	x_1	x_2	x_3	…	x_m
n_i	n_1	n_2	n_3	…	n_m

1) *Выборочное среднее* (обозначается $\bar{x}$) – это среднее арифметическое всех значений выборки:

$$\bar{x} = \frac{1}{n} \cdot \sum_{i=1}^{m} x_i \cdot n_i \quad \text{или} \quad \bar{x} = \sum_{i=1}^{m} x_i \cdot \omega_i \,.$$

Выборочное среднее показывает среднее значение, вокруг которого группируются варианты. $\bar{x}$ является *несмещённой* и *состоятельной* оценкой математического ожидания.

Свойства средней арифметической $\bar{x}$:

1. $\overline{C} = C$, где С – постоянная;
2. $\overline{Cx} = C \cdot \bar{x}$;
3. $\overline{x + C} = \bar{x} + C$;
4. $\overline{x - \bar{x}} = 0$;
5. $\overline{x + y} = \bar{x} + \bar{y}$.

2) *Выборочная дисперсия* – это среднее арифметическое квадратов отклонений значений выборки от выборочной средней.

$$S^2 = \frac{1}{n}\sum_{i=1}^{m}(x_i - \bar{x})^2 \cdot n_i .$$

Выборочная дисперсия показывает меру разброса данных вокруг среднего значения. Она является *состоятельной* и *смещённой* оценкой *дисперсии*.

Свойства средней арифметической $\bar{x}$:

1. $S^2(C) = 0$, где С – постоянная;
2. $S^2(C \cdot x) = C^2 \cdot S^2(x)$;
3. $S^2(x + C) = S^2(x)$;
4. $S^2 = \overline{x^2} - (\bar{x})^2$, где $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{m} x_i^{\,2} \cdot n_i$

Алгоритм расчёта выборочной дисперсии

1) Вычисляется объём выборки n;

2) Находится выборочная средняя $\bar{x}$;

3) Определяется величина $\overline{x^2} = \frac{1}{n}\sum_{i=1}^{m} x_i^{\,2} \cdot n_i$;

4) $S^2 = \overline{x^2} - (\bar{x})^2$.

Для оценки степени отклонения от среднего значения удобно иметь дело с величиной той же размерности, что и величина X. Для этого вводится понятие выборочного среднеквадратического отклонения.

3) *Выборочное среднеквадратической отклонение* – это арифметическое значение корня квадратного из его дисперсии, т.е.

$$S = \sqrt{S^2} .$$

4) *Размах выборки* (обозначается R) – это разность между максимальной и минимальной вариантами или длина интервала, которому принадлежат все варианты выборки:

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

Для подробного описания особенностей распределения используют дополнительные характеристики – моменты распределения. Выяснение общего характера предполагает не только оценку степени его однородности, но и позволяет исследовать форму распределения.

Средняя из *k*-х степеней отклонений вариант x_i от некоторой постоянной величины A называется *моментом k-го порядка*:

$$M_k = \overline{(x_i - A)^k} .$$

При расчёте средних в качестве весов можно использовать частоты, относительные частоты или вероятности. При использовании в качестве весов частот или относительных частот моменты называются *эмпирическими* (обозначаются $\tilde{M}_k$), а при использовании вероятностей – *теоретическими* (обозначаются M_k). Порядок момента определяется величиной *k*.

Эмпирический момент k-го порядка – это отношение суммы произведений *k*-х степеней отклонений вариант от постоянной величины A на частоты к объёму выборки:

$$\tilde{M}_k = \frac{1}{n}\sum_{i=1}^{m}(x_i - A)^k \cdot n_i .$$

Практически используют моменты первых четырех порядков. Если $A = 0$, то моменты *начальные;* $A = \bar{x}$, то моменты *центральные*; A – произвольное число, то моменты *условные*.

Нормальное распределение является одним из самых распространённых в применениях математической статистики. Для оценки отклонения эмпирического распределения от нормального используются *нормированные* моменты.

Порядок момента	Начальные моменты	Центральные моменты	Нормированные моменты
$k=0$	$\tilde{\nu}_0 = 1$	$\tilde{\mu}_0 = 1$	$\tilde{m}_0 = 1$
$k=1$	$\tilde{\nu}_1 = \bar{x}$	$\tilde{\mu}_1 = 0$	$\tilde{m}_1 = 0$
$k=2$	$\tilde{\nu}_2 = \frac{1}{n}\sum_{i=1}^{m} x_i^{\,2} \cdot n_i$	$\tilde{\mu}_2 = S^2$	$\tilde{m}_2 = 1$
$k=3$	$\tilde{\nu}_3 = \frac{1}{n}\sum_{i=1}^{m} x_i^{\,3} \cdot n_i$	$\tilde{\mu}_3 = \frac{1}{n}\sum_{i=1}^{m}(x_i - \bar{x})^3 \cdot n_i$	$\tilde{m}_3 = \tilde{A}_S = \frac{\tilde{\mu}_3}{S^3}$
$k=4$	$\tilde{\nu}_4 = \frac{1}{n}\sum_{i=1}^{m} x_i^{\,4} \cdot n_i$	$\tilde{\mu}_4 = \frac{1}{n}\sum_{i=1}^{m}(x_i - \bar{x})^4 \cdot n_i$	$\tilde{m}_4 = \frac{\tilde{\mu}_4}{S^4}$
Общие формулы			
k	$\tilde{\nu}_k = \frac{1}{n}\sum_{i=1}^{m} x_i^{\,k} \cdot n_i$	$\tilde{\mu}_k = \frac{1}{n}\sum_{i=1}^{m}(x_i - \bar{x})^k \cdot n_i$	$\tilde{m}_k = \frac{\tilde{\mu}_k}{S^k}$

Для облегчения расчётов можно воспользоваться соотношениями между начальными и центральными моментами:

$$\tilde{\mu}_2 = \tilde{\nu}_2 - \tilde{\nu}_1^2,$$

$$\tilde{\mu}_3 = \tilde{\nu}_3 - 3\tilde{\nu}_1 \cdot \tilde{\nu}_2 + 2\tilde{\nu}_1^3,$$

$$\tilde{\mu}_3 = \tilde{\nu}_4 - 4\tilde{\nu}_1 \cdot \tilde{\nu}_3 + 6\tilde{\nu}_1^2 \cdot \tilde{\nu}_2 - 3\tilde{\nu}_1^4.$$

Из математической статистики известно, что при увеличении объёма статистической совокупности $(m \to \infty)$ и одновременного уменьшении интервала группировки полигон либо гистограмма распределения всё более и более приближается к некоторой плавной кривой, являющейся для указанных графиков пределом. Эта кривая называется *эмпирической кривой распределения* и представляет собой графическое изображение в виде непрерывной линии изменения частот, функционально связанного с изменением вариант. В статистике различают следующие виды кривых распределения: одновершинные и многовершинные кривые.

Однородные совокупности описываются одновершинными распределениями. Многовершинность распределения свидетельствует о неоднородности изучаемой совокупности или о некачественном выполнении группировки. Одновершинные кривые распределения делятся на симметричные, умеренно асимметричные и крайне асимметричные.

Распределение называется *симметричным*, если частоты любых 2-х вариантов, равноотстоящих в обе стороны от центра распределения, равны между собой. Для характеристики асимметрии используют коэффициент асимметрии.

При симметричном распределении варианты, равноудаленные от $\bar{x}$, имеют одинаковую частоту, поэтому каждый центральный момент нечётной степени равен 0. Для несимметричного распределения – не равен 0, следовательно, любой из этих моментов может служить для оценки симметрии. Поэтому выбран нечётный момент наименьшего порядка, не равный нулю или μ_3. Чтобы получить безразмерную характеристику, его делят на S^3 (так как μ_3 имеет размерность куба рассматриваемой случайной величины).

5) *Коэффициент асимметрии* $\tilde{A}_S$ – это отношение центрального момента третьего порядка к кубу выборочного среднеквадратического отклонения:

$$\tilde{A}_S = \frac{\tilde{\mu}_3}{S^3}.$$

В одновершинных распределениях величина этого показателя изменяется от –1 до +1. в симметричных распределениях $\tilde{A}_S = 0$.

Если $\tilde{A}_S < 0$, то имеет место левосторонняя асимметрия, а если $\tilde{A}_S > 0$ – правосторонняя асимметрия. Чем ближе по модулю $\tilde{A}_S$ к 1, тем асимметрия существеннее:

– если $\left|\tilde{A}_S\right| \le 0{,}25$, то асимметрия считается незначительной,

– если $0{,}25 < \left|\tilde{A}_S\right| \le 0{,}5$, то асимметрия считается умеренной,

– если $\left|\tilde{A}_S\right| > 0{,}5$, то асимметрия значительная.

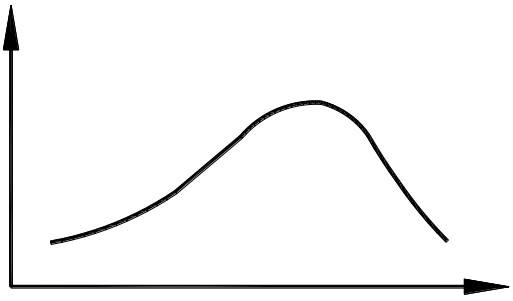

Рис. 5 Правосторонняя асимметрия

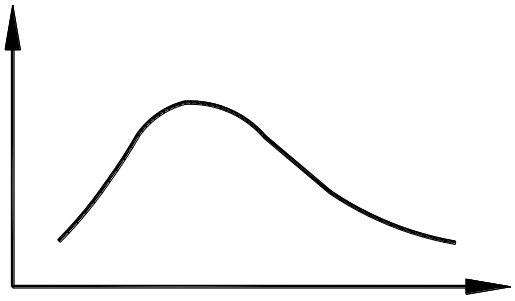

Рис. 6 Левосторонняя асимметрия

6) Для нормального распределения $\mu_4 = 3$, поэтому для оценки крутизны исследуемого распределения в сравнении с нормальным из μ_4 вычитается 3. *Эксцесс* $\tilde{E}_X$ – это уменьшённое на три единицы отношение центрального момента четвёртого порядка к четвёртой степени среднеквадратического отклонения:

$$\tilde{E}_X = \frac{\tilde{\mu}_4}{S^4} - 3.$$

Кривые распределения, у которых $\tilde{E}_X < 0$, менее крутые, имеют более плоскую вершину и называются *плосковершинными*. Кривые распределения, у которых $\tilde{E}_X > 0$, более крутые, имеют острую вершину и называются *островершинными*.

Рассматривая формулы моментов, можно видеть, что начальный момент первого порядка представляет собой среднюю арифметическую и используется как показатель центра распределения. Центральный момент первого порядка (нулевое свойство средней арифметической) всегда равен нулю.

Центральный момент второго порядка представляет собой дисперсию и служит основной мерой колеблемости признака. Центральный момент третьего порядка равен нулю в симметричном распределении и используется для определения показателя асимметрии. Центральный момент четвертого порядка применяется при вычислении показателя эксцесса. Начальные моменты второго, третьего и четвертого порядка, так же как и условные моменты, самостоятельного значения не имеют, а используются для упрощения вычислений центральных моментов.

Пример решения задания 2

Задан вариационный ряд выборки. Найти: выборочное среднее, выборочную дисперсию, выборочное среднеквадратическое отклонение, размах выборки, асимметрию, эксцесс.

x_i	3	5	6	8	9	10	14
n_i	2	10	15	20	38	11	4

Решение

Объём выборки: $n = 2+10+15+20+38+11+4 = 100$.

1) Выборочное среднее:

$$\bar{x} = \frac{3\cdot 2 + 5\cdot 10 + 6\cdot 15 + 8\cdot 20 + 9\cdot 38 + 10\cdot 11 + 14\cdot 4}{100} = \frac{814}{100} = 8{,}14.$$

2) Выборочная дисперсия:

$$\overline{x^2} = \frac{3^2\cdot 2 + 5^2\cdot 10 + 6^2\cdot 15 + 8^2\cdot 20 + 9^2\cdot 38 + 10^2\cdot 11 + 14^2\cdot 4}{100} = \frac{7050}{100} = 70{,}5;$$

$$S^2 = 70{,}5 - 8{,}14^2 \approx 4{,}2.$$

3) Выборочное среднеквадратическое отклонение:

$$S = \sqrt{S^2} = \sqrt{\tilde{\mu}_2} = \sqrt{4{,}2} \approx 2{,}05.$$

4) Размах выборки: $R = 14 - 3 = 11$.

5) Начальные моменты:

– 1-го порядка $\tilde{\nu}_1 = \bar{x} = 8{,}14$;

– 2-го порядка $\tilde{\nu}_2 = \overline{x^2} = 70{,}5$;

– 3-го порядка

$$\tilde{\nu}_3 = \frac{3^3\cdot 2 + 5^3\cdot 10 + 6^3\cdot 15 + 8^3\cdot 20 + 9^3\cdot 38 + 10^3\cdot 11 + 14^3\cdot 4}{100} = \frac{64462}{100} \approx 644{,}6;$$

– 4-го порядка

$$\tilde{\nu}_4 = \frac{3^4 \cdot 2 + 5^4 \cdot 10 + 6^4 \cdot 15 + 8^4 \cdot 20 + 9^4 \cdot 38 + 10^4 \cdot 11 + 14^4 \cdot 4}{100} = \frac{620754}{100} \approx 6207{,}5.$$

6) Центральные моменты

– 2-го порядка $\tilde{\mu}_2 = S^2 = 4{,}2$;

– 3-го порядка

$$\tilde{\mu}_3 = \tilde{\nu}_3 - 3\tilde{\nu}_1 \cdot \tilde{\nu}_2 + 2\tilde{\nu}_1^3 = 644{,}6 - 3 \cdot 8{,}14 \cdot 70{,}5 + 2 \cdot 8{,}14^3 \approx 1{,}7;$$

– 4-го порядка

$$\tilde{\mu}_4 = \tilde{\nu}_4 - 4\tilde{\nu}_1 \cdot \tilde{\nu}_3 + 6\tilde{\nu}_1^2 \cdot \tilde{\nu}_2 - 3\tilde{\nu}_1^4 = 6207{,}5 - 4 \cdot 8{,}14 \cdot 644{,}6 + 6 \cdot 8{,}14^2 \cdot 70{,}5 - 3 \cdot 8{,}14^4 \approx 76{,}1.$$

7) Асимметрия $\tilde{A}_S = \frac{\tilde{\mu}_3}{S^3} = \frac{1{,}7}{2{,}05^3} \approx 0{,}2$.

$0{,}25 < \left|\tilde{A}_S\right| \le 0{,}5$, то есть асимметрия умеренная.

8) Эксцесс $\tilde{E}_X = \frac{\tilde{\mu}_4}{S^4} - 3 = \frac{76{,}1}{2{,}05^4} - 3 \approx 4{,}3 - 3 = 0{,}3$.

Так как $\tilde{E}_X > 0$, то кривая распределения является островершинной.

3. Интервальные оценки параметров распределения

До этого мы рассматривали вопросы об оценке параметра Θ одним числом. Такая оценка называлась точечной. В ряде задач требуется найти для неизвестного параметра не только подходящее значение, но и оценить его точность и надёжность.

Доверительным интервалом для параметра Θ называется интервал $(\varepsilon_1; \varepsilon_2)$, в котором с заранее заданной вероятностью содержится истинное значение этого параметра.

$$P(\varepsilon_1 < \Theta < \varepsilon_2) = \gamma,$$

где γ – доверительная вероятность или надёжность.

Смысл доверительного интервала состоит в том, что при многократном повторении выборки объёма n в относительной доле случаев, равной γ, доверительный интервал, соответствующий доверительной вероятности γ, накрывает истинное значение оцениваемого параметра. Таким образом, чем больше γ, тем вероятнее, что реализация доверительного интервала содержит неизвестный параметр. Однако с ростом доверительной

вероятности γ в среднем растёт длина доверительного интервала, то есть уменьшается точность доверительного оценивания. Выбор доверительной вероятности определяется конкретными условиями, обычно используются значения γ, равные 0,90; 0,95; 0,99; 0,999.

Вероятность

$$\alpha = 1 - \gamma$$

называется *уровнем значимости* и характеризует относительное число ошибочных заключений в общем числе заключений.

Пусть для параметра Θ на основании выборочных данных получена несмещённая оценка $\tilde{\Theta}$. Найдём такое положительное значение ε, для которого событие $\left|\Theta - \tilde{\Theta}\right| < \varepsilon$ с вероятностью γ можно считать достоверным.

$$\left|\Theta - \tilde{\Theta}\right| < \varepsilon,$$

$$-\varepsilon < \Theta - \tilde{\Theta} < \varepsilon,$$

$$\tilde{\Theta} - \varepsilon < \Theta < \tilde{\Theta} + \varepsilon.$$

Интервал $\left(\tilde{\Theta} - \varepsilon; \tilde{\Theta} + \varepsilon\right)$ называется *доверительным*, ε *– погрешность (точность)* доверительного интервала, $\varepsilon < \tilde{\Theta}$.

Для доверительного интервала вида $(a;b)$ справедливы следующие соотношения:

$$\varepsilon = \frac{b-a}{2},$$

$$\tilde{\Theta} = \frac{a+b}{2}.$$

Общая схема построения доверительных интервалов

1. Из генеральной совокупности случайной величины X с известным распределением $f(x,\Theta)$ извлекается выборка объёма n, по которой находится точечная оценка $\tilde{\Theta}$ параметра Θ.

2. Строится новая случайная величина $Y(\Theta)$, связанная с параметром и имеющая известную плотность вероятности $f(y,\Theta)$. Построение случайной величины $Y(\Theta)$ и подбор соответствующего (или близкого) типа распределения для неё определяется свойствами точечной оценки $\tilde{\Theta}$ (как случайной величины).

3. Задаётся уровень значимости α $(\alpha = 0{,}1;\ 0{,}05;\ 0{,}01;\ 0{,}001)$, что соответствует надёжности $\gamma = 1 - \alpha$ $(\gamma = 0{,}9;\ 0{,}95;\ 0{,}99;\ 0{,}999)$.

4. Используя плотность распределения $f(y,\Theta)$ случайной величины $Y(\Theta)$, определяются два числа C_1 и C_2 так, чтобы $P(C_1 < Y(\Theta) < C_2) = \int_{C_1}^{C_2} f(y,\Theta)dy = 1-\alpha$. Значения C_1 и C_2 определяются, как правило, из условий $P(Y(\Theta) \le C_1) = P(Y(\Theta) \ge C_2) = \frac{\alpha}{2}$.

Эти значения определяются по таблицам как квантили распределения случайной величины $Y(\Theta)$. Используя связь случайных величин $Y(\Theta)$ и $\Theta - \tilde{\Theta}$, неравенство $C_1 < Y(\Theta) < C_2$ преобразуют в равносильное неравенство $-\varepsilon < \Theta - \tilde{\Theta} < \varepsilon$ такое, что $P(-\varepsilon < \Theta - \tilde{\Theta} < \varepsilon) = 1-\alpha = \gamma$. Полученный интервал $(\tilde{\Theta} - \varepsilon; \tilde{\Theta} + \varepsilon)$, содержащий неизвестный параметр Θ с вероятностью γ, является интервальной оценкой параметра Θ. Положительное число ε характеризует точность оценки.

4. Доверительный интервал для оценки математического ожидания нормально распределённой случайной величины при известном среднеквадратическом отклонении

Пусть количественный признак X генеральной совокупности распределён нормально, среднеквадратическое отклонение σ известно. Требуется оценить неизвестное математическое ожидание a по выборочной средней $\bar{x}$.

В данном случае в качестве случайной величины $Y(\Theta)$ берётся величина $Y(\Theta) = \frac{\bar{X} - \bar{x}}{\frac{\sigma}{\sqrt{n}}}$, которая при достаточно больших объёмах выборки приближённо распределена по нормальному закону $N(0,1)$. Поэтому с заданной надёжностью γ доверительный интервал имеет вид $\left(\bar{x} - \frac{t \cdot \sigma}{\sqrt{n}}; \bar{x} + \frac{t \cdot \sigma}{\sqrt{n}}\right)$.

Таким образом, если исследуемая случайная величина распределена по нормальному закону с известным среднеквадратическим отклонением σ, то доверительный интервал для математического ожидания определяется неравенством:

$$\bar{x} - \frac{t \cdot \sigma}{\sqrt{n}} < a < \bar{x} + \frac{t \cdot \sigma}{\sqrt{n}},$$

где $\tilde{\Theta} = \bar{x}$ – точечная оценка математического ожидания ($\bar{x}$ – выборочное среднее);

$\varepsilon = \frac{t \cdot \sigma}{\sqrt{n}}$ – точность оценки;

n – объём выборки;

t – квантиль нормального распределения или значение аргумента функции Лапласа (приложение 2 [1, 2]), при котором $2\Phi(t) = \gamma \Rightarrow \Phi(t) = \frac{\gamma}{2}$.

Пример решения задания 3

Найти доверительный интервал для оценки математического ожидания a нормального распределения с надёжностью $\gamma = 0{,}95$, зная выборочное среднее $\bar{x} = 2{,}3$, объём выборки $n = 49$ и генеральное среднеквадратическое отклонение $\sigma = 1{,}4$.

Решение

Воспользуемся формулой: $\Phi(t) = \frac{\gamma}{2} = \frac{0{,}95}{2} = 0{,}475$, далее по таблице приложения 2 [1, 2] находим $t = 1{,}96$. Искомый доверительный интервал:

$$2{,}3 - \frac{1{,}96 \cdot 1{,}4}{\sqrt{49}} < a < 2{,}3 + \frac{1{,}96 \cdot 1{,}4}{\sqrt{49}} \text{ или } 1{,}908 < a < 2{,}692.$$

Ответ: $1{,}908 < a < 2{,}692$.

Смысл полученного результата: если произведено достаточно большое количество выборок по 49 элементов в каждой, то 95% из них определяют такие доверительные интервалы, в которых a заключено, и лишь в 5% случаев значение a может выйти за границы доверительного интервала.

5. Доверительный интервал для оценки математического ожидания нормально распределённой случайной величины при неизвестном среднеквадратическом отклонении

Пусть количественный признак X генеральной совокупности распределён нормально, причём среднеквадратическое отклонение σ неизвестно. Требуется оценить неизвестное математическое ожидание a с помощью доверительного интервала с заданной точностью γ.

Известно, что если случайная величина Z распределена нормально по закону $N(0,1)$, а величина V имеет χ^2 – распределение с $k = n-1$ степенью свободы, причём эти величины независимы, то случайная величина $T = \dfrac{Z}{\sqrt{\dfrac{V}{n}}}$ имеет t-распределение Стьюдента с $k = n-1$ степенью свободы.

В частности, такими свойствами обладают случайные величины $Z = (\overline{X} - a)\dfrac{\sqrt{n}}{\sigma}$, $V = (n-1)\dfrac{{S_X}^{*2}}{\sigma^2}$. Таким образом, мы можем использовать в качестве $Y(\Theta)$ случайную величину $T = \dfrac{\overline{X} - a}{\dfrac{S_X^*}{\sqrt{n}}}$, которая имеет распределение Стьюдента с $k = n-1$ степенями свободы. Здесь $\overline{X}$ – выборочная средняя, S_X^* – исправленное выборочное среднеквадратическое отклонение, n – объём выборки.

По таблице t-распределения Стьюдента по заданным значениям n и γ находится квантиль t_γ, удовлетворяющий условию

$$P(-t_\gamma < T < t_\gamma) = P\left(\overline{X} - \frac{t_\gamma \cdot S_X^*}{\sqrt{n}} < a < \overline{X} + \frac{t_\gamma \cdot S_X^*}{\sqrt{n}}\right) = \gamma.$$

Таким образом, доверительный интервал имеет вид $\left(\bar{x} - \dfrac{t_\gamma \cdot S^*}{\sqrt{n}}; \bar{x} + \dfrac{t_\gamma \cdot S^*}{\sqrt{n}}\right)$. Он содержит неизвестный параметр a с надёжностью γ. При построении случайные величины $\overline{X}$ и S_X^* заменяются неслучайными значениями $\bar{x}$ и S^*, найденными по

данной выборке. По таблице по заданным значениям n и γ можно найти t_γ.

Таким образом, если среднеквадратическое отклонение нормально распределённой случайной величины неизвестно, то доверительный интервал для оценки математического ожидания определяется соотношением:

$$\bar{x} - \frac{t_\gamma \cdot S^*}{\sqrt{n}} < a < \bar{x} + \frac{t_\gamma \cdot S^*}{\sqrt{n}},$$

где $\tilde{\Theta} = \bar{x}$ – точечная оценка математического ожидания ($\bar{x}$ – выборочное среднее);

$\varepsilon = \frac{t_\gamma \cdot S^*}{\sqrt{n}}$ – точность оценки;

S^* – исправленное выборочное среднеквадратическое отклонение;

n – объём выборки;

t_γ – квантиль распределения Стьюдента, определяется:

а) по таблице приложения 3 [1, 2] в зависимости от объёма n и надёжности γ

или

б) по таблице приложения 6 [1, 2] в зависимости от числа степеней свободы $k = n - 1$ и уровня значимости $\alpha = 1 - \gamma$.

Пример решения задания 4

Найти доверительный интервал для оценки математического ожидания a нормального распределения с надёжностью $\gamma = 0{,}95$, зная выборочное среднее $\bar{x} = 8$, объём выборки $n = 10$ и исправленную выборочную дисперсию $S^{*2} = 5{,}76$.

Решение

Найдём исправленное среднеквадратическое отклонение $S^* = \sqrt{5{,}76} = 2{,}4$.

Квантиль распределения Стьюдента определим двумя способами:

а) $n = 10$, $\gamma = 0{,}95 \Rightarrow t_\gamma = 2{,}26$

или

б) $k = 10 - 1 = 9$, $\alpha = 1 - 0{,}95 = 0{,}05 \Rightarrow t_\gamma = 2{,}26$.

Искомый доверительный интервал $8-\frac{2,26\cdot 2,4}{\sqrt{10}}<a<8+\frac{2,26\cdot 2,4}{\sqrt{10}}$ или $6,3<a<9,7$.

Ответ: $6,3<a<9,7$.

6. Проверка статистических гипотез

Если закон распределения неизвестен, то выдвигают гипотезу о его виде. Возможен также случай, когда закон распределения известен, а его параметр Θ неизвестен. Тогда есть основание предположить, что неизвестный параметр Θ равен определённому значению Θ_0 и выдвигают гипотезу $\Theta=\Theta_0$.

Статистической называют гипотезу о виде неизвестного распределения или о параметрах известных распределений. гипотезы бывают простые и сложные.

Простой называют гипотезу, содержащую только одно предположение. Сложная гипотеза состоит из конечного числа простых гипотез.

Выдвинутая гипотеза называется *нулевой* и обозначается H_0, а гипотеза, противоречащая нулевой – *конкурирующей* и обозначается H_1.

Выдвинутая гипотеза может быть правильной или неправильной, поэтому возникает необходимость её проверки. В результате проверки могут быть допущены ошибки двух видов.

Ошибка первого рода состоит в том, что будет отвергнута правильная гипотеза, *ошибка второго рода* – будет принята неправильная гипотеза.

Решение признать верной гипотезу H_0 или H_1 принимается по значению некоторой функции выборки, называемой *статистическим критерием*.

Значение критерия, вычисленное по выборке, называется *наблюдаемым* и обозначается $k_{набл}$. Множество значений критерия можно разделить на два непересекающихся подмножества: подмножество значений критерия, при которых гипотеза H_0 принимается, называется *допустимой областью;* подмножество

значений критерия, при которых гипотеза H_0 отвергается и принимается гипотеза H_1, называется *критической областью*.

Критическими точками называются точки, отделяющие критическую область от допустимой. Эти точки являются *табличными* или *критическими* значениями критерия и обозначаются $k_{крит}$.

При проверке гипотез следует по возможности уменьшить вероятности принятия неправильных решений. Допустимая вероятность ошибки I рода $\alpha = 1 - \gamma$ называется *уровнем значимости.*

Для определения критической области используют уровень значимости и учитывают вид альтернативной гипотезы H_1.

$k_{крит}$ определяют по таблицам распределения данного критерия если $k_{набл} \le k_{крит}$, то гипотеза H_0 принимается, если $k_{набл} > k_{крит}$, то принимается гипотеза H_1.

Пример решения задания 5

Для заданного интервального ряда выборки проверить гипотезу: закон распределения генеральной совокупности является нормальным.

Интервал	$[2,2;3,0)$	$[3,0;3,8)$	$[3,8;4,6)$	$[4,6;5,4)$	$[5,4;6,2)$	$[6,2;7,0)$	$[7,0;7,8]$
n_i	5	10	35	20	15	8	7

Решение

Выдвигаются нулевая и конкурирующая гипотезы:

H_0: закон распределения генеральной совокупности является нормальным;

H_1: генеральная совокупность имеет закон распределения отличный от нормального.

Интервальный вариационный ряд преобразуется в дискретный. Для этого интервалы заменяются соответствующими им серединами, а частоты остаются прежними.

x_i	2,6	3,4	4,2	5,0	5,8	6,6	7,4
n_i	5	10	35	20	15	8	7

По полученным данным находятся выборочное среднее и выборочное среднеквадратическое отклонение.

1) $n = 5 + 10 + 35 + 20 + 15 + 8 + 7 = 100$;

2) $\bar{x} = \dfrac{2{,}6 \cdot 5 + 3{,}4 \cdot 10 + 4{,}2 \cdot 35 + 5{,}0 \cdot 20 + 5{,}8 \cdot 15 + 6{,}6 \cdot 8 + 7{,}4 \cdot 7}{100} \approx 4{,}8$;

3) $\overline{x^2} = \dfrac{2{,}6^2 \cdot 5 + 3{,}4^2 \cdot 10 + 4{,}2^2 \cdot 35 + 5{,}0^3 \cdot 20 + 5{,}8^2 \cdot 15 + 6{,}6^2 \cdot 8 + 7{,}4^2 \cdot 7}{100} \approx 25$;

4) $S^2 = 25 - 4{,}8^2 = 1{,}96$;

5) $S = \sqrt{1{,}96} \approx 1{,}4$.

Гипотеза проверяется с помощью случайной величины $\chi^2 = \sum_{i=1}^{m} \dfrac{\left(n_i - n_i''\right)^2}{n_i''}$, число степеней свободы которой находится по формуле:

$$k = l - r - 1,$$

где l – число интервалов, на которые разбит вариационный ряд;

r – число параметров распределения, которые оценены по данным выборки (для нормального распределения $r = 2$, для распределения Пуассона $r = 1$).

Значит $k = 7 - 1 - 1 = 5$. Задаётся уровень значимости $\alpha = 0{,}05$.

По уровню значимости и числу степеней свободы критическая точка правосторонней критической области $(0{,}05; 5)$ находится из приложения 5 [1, 2] и равна $\chi^2_{крит} = 11{,}1$.

Предварительно определим теоретические частоты по формуле $n_i'' = n \cdot \left(\Phi\left(\dfrac{x_i - \bar{x}}{S} \right) - \Phi\left(\dfrac{x_{i-1} - \bar{x}}{S} \right) \right)$.

Расчёты представлены в таблице.

x_i	3,0	3,8	4,6	5,4	6,2	7,0	7,8
$\frac{x_i - \bar{x}}{S}$	–1,28	–0,71	–0,14	0,43	1,00	1,57	2,14
$\Phi\left(\frac{x_i - \bar{x}}{S}\right)$	–0,3997	–0,2611	–0,0557	0,1664	0,3413	0,4418	0,4838
x_{i-1}	2,2	3,0	3,8	4,6	5,4	6,2	7,0
$\frac{x_{i-1} - \bar{x}}{S}$	–1,86	–1,28	–0,71	–0,14	0,43	1,00	1,57
$\Phi\left(\frac{x_{i-1} - \bar{x}}{S}\right)$	–0,4686	–0,3997	–0,2611	–0,0557	0,1664	0,3413	0,4418
$\Phi\left(\frac{x_i - \bar{x}}{S}\right) - \Phi\left(\frac{x_{i-1} - \bar{x}}{S}\right)$	$\approx 0,07$	$\approx 0,13$	$\approx 0,21$	$\approx 0,23$	$\approx 0,18$	$\approx 0,13$	$\approx 0,05$
n_i''	7	13	21	23	18	13	5
$\frac{\left(n_i - n_i''\right)^2}{n_i''}$	0,571	0,692	9,333	0,391	0,500	1,923	0,800

$$\chi^2_{набл} = 0,571 + 0,692 + 9,333 + 0,391 + 0,500 + 1,923 + 0,800 = 14,21$$

По уровню значимости и числу степеней свободы критическая точка правосторонней критической области $(0,05;5)$ находится из приложения 5 [1, 2] и равна $\chi^2_{крит} = 11,1$.

Так как $\chi^2_{набл} > \chi^2_{крит}$, то гипотеза H_0 о нормальном распределении отвергается.

Ответ: закон распределения генеральной совокупности не является нормальным.

Список рекомендуемой литературы

1. Гмурман В.Е. Теория вероятностей и математическая статистика [Текст]: учебное пособие. -М.: ЮРАЙТ, 2012.–479с.
2. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике [Текст]: учебное пособие. -М.: ЮРАЙТ, 2011.-404с.
3. Бойцова Е.А. Практикум по математике. Спецглавы [Текст]: учебное пособие / Е.А.Бойцова. − Старый Оскол: ТНТ, 2014. – 156с.