Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Приморская государственная сельскохозяйственная академия»
Институт лесного и лесопаркового хозяйства

Кафедра философии и
социально-гуманитарных
дисциплин

Жуплей И.В., Савельева Е.В.

ОСНОВЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ ДЛЯ ЭКОНОМИЧЕСКОГО БАКАЛАВРИАТА

Электронное издание

Уссурийск 2017

УДК 31.
ББК 65.01

Рецензенты: Т.И. Еременко, ведущий специалист-эксперт отдела государственной статистики г. Владивостоке (г. Уссурийск) Приморскстата;

Л. Д. Ердакова, канд. пед. наук, доцент, доцент кафедры физики и теоретической механики ПримИЖТ филиала ДВГУПС в г. Уссурийске, г. Уссурийск

Жуплей, И.В. Основы математической статистики для экономического бакалавриата: учеб. пособие [Электронный ресурс] / И.В. Жуплей, Е.В. Савельева – Электрон. текст. дан. - Уссурийск, 2017. – 153 с. – Режим доступа: http://de.primacad.ru

Учебное пособие включает перечень вопросов, необходимый теоретический минимум, решение основных типов задач, значительный подбор задач и тесты, что позволит эффективно использовать пособие для самостоятельной работы, а также при проведении контролирующих мероприятий по дисциплине (модулю).

Для обучающихся очной и заочной форм обучения по направлению подготовки 38.03.01 Экономика (уровень бакалавриата).

Электронное издание

Издается по решению методического совета ФГБОУ ВО Приморская ГСХА

ПРЕДИСЛОВИЕ

Умение адекватно анализировать экономические процессы – одно из важнейших умений, которыми должен обладать выпускник экономического бакалавриата. А анализ экономических данных требует применения особых приемов, которые позволяли бы обобщать и сопоставлять их, чтобы делать необходимые для науки и практики выводы. Эти приемы изучаются статистикой, основы которой заложены в статистике математической. В то же время сейчас все большее внимание уделяется самостоятельной работе обучающихся, повышению качества специалистов.

В связи с чем для подготовки экономистов необходимы не только учебники и справочники по тем или иным разделам статистики, в том числе и математической, но и пособия, направленные для самостоятельного изучения отдельных тем и выработки навыков по решению обязательного набора задач в рамках каждой темы.

Предлагаемое пособие составлено в соответствии с Государственным образовательным стандартом высшего образования для экономических специальностей на основе дидактического материала, разработанных авторами для обучающихся по направлению подготовки 38.03.01 Экономика.

Многим задачам дана современная экономическая интерпретация, при их составлении использовались фактические данные экономики сельского хозяйства. Основные понятия также иллюстрируются примерами экономического содержания.

Пособие состоит из пяти глав («Выборка и ее представление», «Числовые характеристики выборки», «Интервальные оценки числовых характеристик генеральной совокупности», «Статистические гипотезы», «Элементы корреляционного анализа»), в каждой из которых содержится необходимый теоретический минимум, примеры решения типовых задач, подбор заданий для самостоятельного решения, тестовые задания и перечень вопросов для самопроверки.

В пособии не проводится разбиение учебных тем «по часам», так как это целесообразно решать каждому преподавателю индивидуально с учётом своего календарного плана, уровня подготовленности учебной группы, специфики обучения (очная или заочная форма обучения) и других факторов. Поэтому количество задач по каждой теме рассчитано на максимальное число часов по предмету в рамках типовых программ.

Задания расположены преимущественно по возрастанию их степени сложности и сгруппированы следующим образом: сначала предложены задачи обязательного уровня (желательно их решение в аудитории), затем (после черты) – задачи, которые можно предложить студентам для самостоятельного решения (также стандартный набор заданий), и третья группа (после двойной черты) – задачи для дополнительного решения. Задачи повышенной трудности обозначены значком ■.

В пособии сделана попытка активизации самостоятельной работы студентов посредством введения в текст изложения решения задач значков:

$\boxed{?}$ – означает, что обучающемуся предлагается самостоятельно записать номер формулы, таблицы и т.п. или требуется выполнить вычисления или алгебраические преобразования, а затем сверить с ответом; $\boxed{!}$ – акцентирование внимания.

Предназначено пособие в первую очередь для организации самостоятельной работы обучающихся сельскохозяйственных вузов, обучающихся направлению подготовки 38.03.01 Экономика, но может быть использовано и для других экономических специальностей (не обязательно сельскохозяйственной направленности).

Некоторые *обозначения*, используемые в пособии:

• – Замечание; $\sum\limits_{i=m}^{n}$ – суммирование;

$\prod\limits_{i=1}^{m}$ – произведение; $\overline{1,n} = 1,2,3,..,n$.

ГЛАВА 1. ВЫБОРКА И ЕЕ ПРЕДСТАВЛЕНИЕ

1.1. Необходимый теоретический минимум

Математической статистикой называется наука, занимающаяся разработкой методов получения, описания и обработки опытных данных с целью изучения закономерностей случайных массовых явлений.

Понятие вариационного ряда. Виды вариационных рядов

Определение 1.1. ***Генеральной совокупностью*** называется совокупность всех возможных однородных предметов или явлений, над которыми проводятся наблюдения, или совокупность всех возможных наблюдений, проводимых над некоторой случайной величиной в одинаковых условиях.

• На практике сплошное обследование применяется сравнительно редко (если совокупность содержит очень большое число элементов). В таких случаях случайным образом выбирают из всей совокупности ограниченное число объектов и подвергают их изучению.

Определение 1.2. ***Выборочной совокупностью (выборкой)*** называется совокупность предметов или явлений, отобранная из соответствующей генеральной совокупности.

Определение 1.3. ***Объемом*** совокупности (генеральной или выборочной) называется общее число ее элементов.

- Если N и n – соответственно объемы генеральной и выборочной совокупности, то, очевидно, что $N>n$.
- *Выборочный метод* заключается в том, что из генеральной совокупности берется выборка значительно меньшего объема и определяются характеристики выборки, которые принимаются в качестве приближенных значений соответствующих характеристик генеральной совокупности.

Выборки бывают:

а) ***повторные*** – если отобранный объект (перед отбором следующего) возвращается в генеральную совокупность;

б) ***бесповторные*** – если отобранный объект не возвращается в генеральную совокупность.

Пусть из генеральной совокупности извлечена выборка X объемом n. Случайный *выбор элемента* рассматривается как независимое наблюдение над величиной ξ, имеющей некоторое распределение вероятностей. Если те значения, которые приняла случайная величина ξ в n наблюдениях, записать не в порядке получения, а в порядке возрастания (то есть ***ранжируя***), то получим упорядоченную выборку $x_1, x_2, \ldots, x_n$,, называемую ***вариационным рядом***.

- Выборка и вариационный ряд несут одну и ту же информацию, но с вариационным рядом работать легче в силу его упорядоченности.
- Чтобы по данным выборки можно было достаточно точно судить об исследуемом признаке генеральной совокупности, требуется, чтобы выборка правильно представляла пропорции генеральной совокупности, то есть была ***репрезентативной (представительной).***

Определение 1.4. **Вариантой** называется значение x_i случайной величины, соответствующее отдельной группе сгруппированного ряда наблюдаемых данных.

Определение 1.5. ***Размахом R*** вариационного ряда называется разность между его наибольшей $x_{\max}$ и наименьшей $x_{\min}$ вариантами:

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

- Если изучается дискретная случайная величина, то при достаточно большом объеме выборки в ней могут содержаться повторяющиеся значения. Для каждого такого значения можно подсчитать, сколько раз оно встретилось в ряде наблюдений.

Определение 1.6. ***Частотой (весом)*** варианты называется численность отдельной группы сгруппированного ряда наблюдаемых данных.

- Если i – индекс варианты, то m_i – число измеренных значений i-ой варианты.

Данные наблюдений, среди которых много повторяющихся, удобно изображать в виде таблицы (табл. 1.1).

Таблица 1.1

Значения x_i	x_1	x_2	…	x_k
Частоты m_i	m_1	m_2	…	m_k

Пример 1. Выборка оценок десяти студентов следующая:

5; 4; 3; 3; 4; 3; 3; 5; 4; 3; 3.

Расположив эти значения в порядке неубывания, получим следующий ряд: 3; 3; 3; 3; 3; 4; 4; 4; 5; 5. Значения 3; 4; 5, принятые случайной величиной в процессе наблюдения (k =3), являются вариантами, а их количества 5; 3; 2 – весы (частоты) соответствующих вариант. В итоге получим вариационный ряд:

Значения x_i	3	4	5
Частоты m_i	5	3	2

Контроль: $\sum_{i=1}^{3} m_i = 5+3+2=10.$

Определение 1.7. ***Относительной частотой варианты*** x_i называется отношение m_i числа повторения x_i к объему выборки n:

$$\tilde{p}_i = \frac{m_i}{n}. \tag{1.1}$$

- Очевидно, что $n = \sum_{i=1}^{k} m_i.$

По таблице, изображающей вариационный ряд, построим таблицу из двух строк, в верхней строке которой указаны в порядке возрастания варианты x_i, а в нижней – соответствующие им относительные частоты $\tilde{p}_i$ (табл. 1.2). Такая таблица называется *таблицей статистического распределения.*

Таблица 1.2

Значения x_i	x_1	x_2	…	x_k
Относительные частоты $\tilde{p}_i$	$\tilde{p}_1 = \frac{m_1}{n}$	$\tilde{p}_2 = \frac{m_2}{n}$	…	$\tilde{p}_k = \frac{m_k}{n}$

Пример 2. Для примера 1 таблица статистического распределения выборки имеет вид:

Значения x_i	3	4	5
Относительные частоты $\tilde{p}_i$	5/10	3/10	2/10

Контроль: $\sum_{i=1}^{3} \tilde{p}_i = 5/10 + 3/10 + 2/10 = 1.$

Определение 1.8. ***Дискретным вариационным рядом*** распределения (***распределением частот*** или ***относительных частот***) называется ранжированная совокупность вариант x_i с соответствующими им частотами или относительными частотами.

Алгоритм составления дискретного вариационного ряда:

– найти минимальное ($x_{\min}$) и максимальное ($x_{\max}$) значения выборки;

– в первый столбец таблицы записать варианты значений случайной величины (генеральной совокупности), начиная с $x_{\min}$ и заканчивая $x_{\max}$;

– просмотреть по одному все элементы выборки в протоколе наблюдений, и подсчитать количество значений, соответствующих каждой варианте (то есть m_i);

– подсчитать количество элементов выборки (ее объем).

• Если изучается случайная величина, имеющая непрерывное распределение вероятностей, то возможные значения заполняют целый интервал или всю числовую ось. В этом случае, скорее всего, вариационный ряд не будет содержать повторяющихся значений. То же самое может иметь место, если наблюдение производится над дискретной случайной величиной, число возможных значений которой очень велико.

•Для выборки, в которой нет повторяющихся значений, таблица статистического распределения будет иметь вид, представленный в таблице 1.3.

Таблица 1.3

Значения x_i	x_1	x_2	…	x_k
Относительные частоты $\tilde{p}_i$	$\frac{1}{n}$	$\frac{1}{n}$	…	$\frac{1}{n}$

• Таблица статистического распределения вида таблицы 1.3 при большом числе наблюдений не содержит полезной информации. Поэтому прибегают к *группировке* данных. Обычно группировку стараются провести таким образом, чтобы значения, различие которых для практики незначимо, попали в один и тот же интервал, а те, различия которых уже значимы, попали в разные интервалы.

Определение 1.9. ***Интервальным вариационным рядом (интервальным распределением частот*** или ***относительных частот)*** называется упорядоченная последовательность интервалов варьирования случайной величины с соответствующими частотами или относительными частотами попаданий в каждый из них значений случайной величины.

Пример 3. В таблице 1.4 приведен пример интервального вариационного ряда, составленный на основании выборки результатов измерения роста 105 студентов (юношей).

Таблица 1.4

Номер интервала i	Рост студентов (интервалы) $x_i < X \le x_{i+1}$	Частота m_i	Относительная частота $\tilde{p}_i$
1	150 –155	4	0,0381
2	155 – 160	–	–
3	160 – 165	2	0,0190
4	165 – 170	19	0,1810
5	170 – 175	19	0,1810
6	175 – 180	26	0,2476
7	180 – 185	21	0,2000
8	185 – 190	10	0,0953
9	195 – 200	2	0,0190
10	200 – 205	2	0,0190
Σ		105	1,0000

Соответствующие относительные частоты (последний столбец таблицы 1.4) найдены по формуле (1.1), а именно:

$$\tilde{p}_1 = \frac{m_1}{n} = \frac{4}{105} \approx 0{,}0381; \quad \ldots \quad \tilde{p}_{10} = \frac{m_{10}}{n} = \frac{2}{105} = 0{,}0190.$$

Алгоритм составления интервального вариационного ряда:

– найти минимальное (x_{min}) и максимальное (x_{max}) значения выборки;

– найти объем n выборки;

– определить оптимальное число интервалов по *формуле Стерджеса*:

$$k = 1 + 3{,}322 \cdot \lg n. \qquad \textbf{(1.2)}$$

– найти длину интервала по формуле:

$$h = \frac{x_{\max} - x_{\min}}{k}. \qquad \textbf{(1.3)}$$

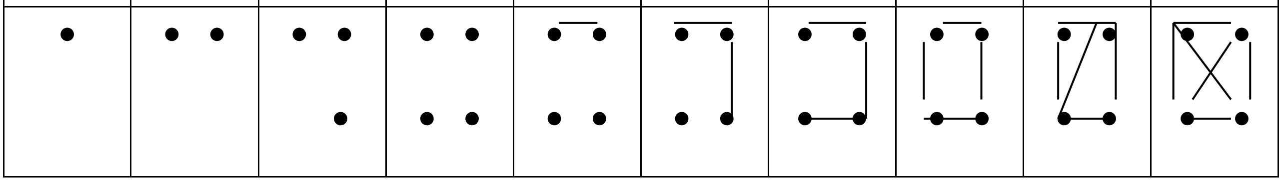
– заполнить первый столбец таблицы интервалами исследуемой совокупности. За начало первого интервала рекомендуется брать величину $$x_{нач} = x_{\min} - 0{,}5 \cdot h. \qquad \textbf{(1.4)}$$

– просмотреть по одному все элементы выборки в протоколе наблюдений, и подсчитать количество значений, соответствующих каждому интервалу (то есть m_i).

• Частоты попадания в интервал можно посчитать следующим образом. Границы интервалов выписывают в столбец. Затем просматривают данные, записанные в том порядке, в котором они были получены. Правее интервала, в который попало значение, ставят точку или черточку. Точки и черточки удобно ставить так, чтобы десять попаданий образовывали «конверт», для каждого интервала подсчитывается число полных «конвертов» и число точек-черточек в неполном (табл 1.5).

Таблица 1.5.

1	2	3	4	5	6	7	8	9	10



Определение 1.10. ***Накопленной частотой*** $m_x^{нак}$ в точке x называют суммарную частоту членов статистической совокупности со значениями признака, меньшими, чем x.

Определение 1.11. ***Накопленной относительной частотой (частостью)*** $p_x^{нак}$ называется отношение накопленной частоты $m_x^{нак}$ к объему выборки n:
$$p_x^{нак} = \frac{m_x^{нак}}{n}. \qquad (1.5)$$

• Каждая генеральная совокупность имеет функцию распределения *F(x)* , которая обычно неизвестна. По выборке можно найти эмпирическую функцию распределения $F^*(x)$, для которой (на основании закона больших чисел) вместо вероятностей p_i берутся относительные частоты p_i/m.

Определение 1.12. ***Эмпирической функцией распределения (выборочной*** или ***функцией распределения выборки)*** называется функция $F^*(x)$, определяющая для каждого значения *x* относительную частоту события $X<x$:

$$F^*(x) = \tilde{P}(X < x) = p_x^{нак}. \qquad \textbf{(1.6)}$$

- Значениями эмпирической функции распределения являются накопленные частости.
- Эмпирическая функция распределения выборки служит для оценки теоретической функции распределения *F(x)* генеральной совокупности.
- Выборочную функцию распределения можно задавать таблично или графически.

Пример 4. Получены следующие данные о распределении 100 рабочих цеха по выработке (в процентах к предыдущему году):

91,8; 97,0; 101,7; 132,5; … ;112,3; 104,2; 141,0; 122,1.

Здесь n=100, x_{min}=97; x_{max}=141. По формуле (1.2) число интервалов k=1+3,322·2=8. Длина интервала (по формуле (1.3)) составит: h=(141 – 97)/8=6. Начало первого интервала найдем по формуле (1.4): $x_{нач} = 97 - 3 = 94.$

• Представленная в таблице 1.6 накопленная частота показывает, сколько наблюдалось вариантов со значением признака, меньшим *x*.

• По найденным в таблице 1.6 накопленным частотам можно ответить на вопросы типа: а) какова доля рабочих, выработка которых по отношению к прошлогодней меньше 100%?; б) какова доля тех, у кого выработка увеличилась в 1,3 раза (то есть на 130%)?; в) чему равна такая выработка, при которой у половины рабочих выработка меньше, а у половины рабочих больше этого значения?

Таблица 1.6

Таблица выработки рабочих (в процентах к предыдущему году)

Номер интервала i	Выработка в процентах (интервалы) $x_i < X \le x_{i+1}$	Количество рабочих (частота) m_i	Доля рабочих (относительная частота) $\tilde{p}_i$	Накопленная частота $m_x^{нак}$	Накопленная относительная (эмпирическая) частота $F^*(x) = p_x^{нак}$
1	94–100	3	0,03	3	0,03
2	100–106	7	0,07	3+7=10	0,10
3	106–112	11	0,11	10+11=21	0,21
4	112–118	20	0,20	21+20=41	0,41
5	118–124	28	0,28	41+28=69	0,69
6	124–130	19	0,19	69+19=88	0,88
7	130–136	10	0,10	88+10=98	0,98
8	136–142	2	0,02	98+2=100	1,00
	Σ	100	1,00	–	–

Ответ.

а) 0,03;

б) 1– 0,88=0,12;

в) примерно 120%.

Графическое представление выборки

Для наглядного представления вариационного ряда (или статистического распределения) пользуются графическим изображением вариационных рядов (полигоном, гистограммой, кумулятой).

Полигон

1. ***Полигон частот*** – это ломаная, отрезки которой соединяют точки $(x_1;m_1),(x_2;m_2), \ldots, (x_k;m_k)$. (См. табл. 1.1). Для построения полигона частот на оси абсцисс откладывают варианты x_i, а на оси ординат – соответствующие им частоты m_i. Точки $(x_i;m_i)$ соединяют отрезками прямых и получают полигон частот.
2. ***Полигон относительных частот*** – это ломаная, отрезки которой соединяют точки $(x_1;\tilde{p}_1),(x_2;\tilde{p}_2), \ldots, (x_k;\tilde{p}_k)$. (См. табл. 1.2). Для построения полигона относительных частот на оси абсцисс откладывают варианты x_i, а на оси ординат – соответствующие им относительные частоты $\tilde{p}_i$. Точки $(x_i;\tilde{p}_i)$ соединяют отрезками прямых и получают полигон частот.

Пример 5. Полигоны частот и относительных частот, построенные на основании таблиц к примерам 1 и 2, показаны соответственно на рис. 1.1 и рис. 1.2.

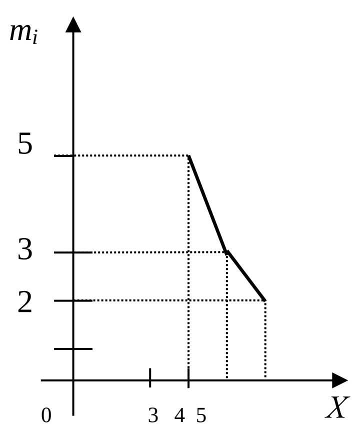


Рис. 1.1.

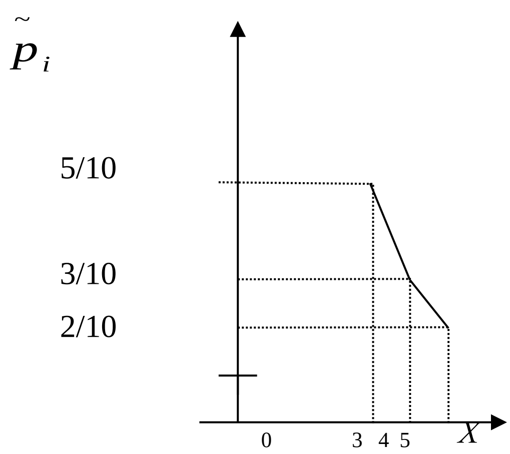


Рис. 1.2

Гистограмма

Гистограмма строится только для интервального вариационного ряда (группированной выборки).

1. ***Гистограмма частот*** – это ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению m_i/h (плотность частоты).

- Площадь гистограммы частот равна сумме всех частот, то есть объему выборки.

2. ***Гистограмма относительных частот*** – это ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению $\tilde{p}_i/h$ (плотность относительной частоты).

- Площадь гистограммы относительных частот равна сумме всех относительных частот, то есть единице.
- Если середины верхних сторон прямоугольников соединить отрезками прямых, а концы этой ломаной еще соединить с серединами соседних интервалов, частоты которых равны 0, а длина равна длине соседнего интервала, то *получим полигон интервального ряда.*

Пример 6. Гистограммы частот и относительных частот, построенные по данным таблицы 1.7, приведены соответственно на рис. 1.3 и рис. 1.4.

Таблица 1.7

i	Интервалы $x_i < X \le x_{i+1}$	Длина интервала h	Частота m_i	Относительная частота $\tilde{p}_i$	Плотность частоты m_i/h	Плотность относительной частоты $\tilde{p}_i/h$
1	2 –5	3	9	9/50	9/3=3	9/150
2	5 –8	3	10	10/50	10/3	10/150
3	8 –11	3	25	25/50	25/3	25/150
4	11 –14	3	6	6/50	6/3=2	6/150
Σ		–	50	1	–	–

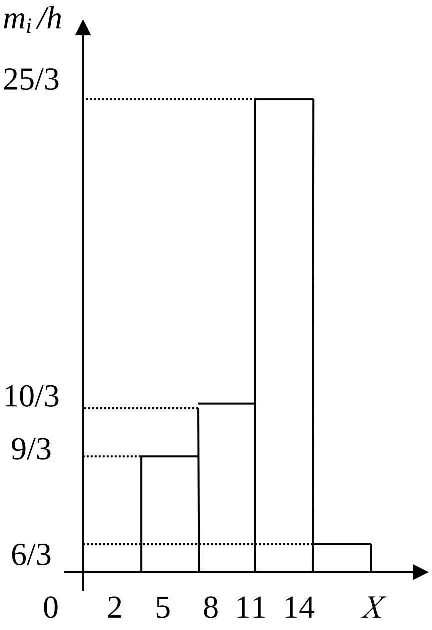


Рис. 1.3

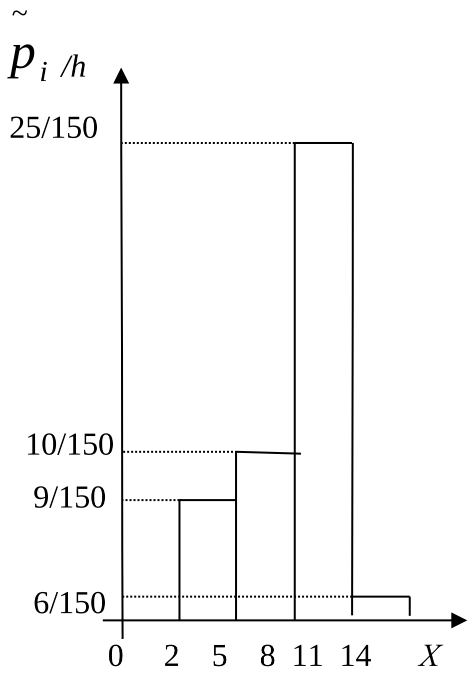


Рис. 1.4

Кумулята

Кумулята (кумулятивная кривая) – график накопленных частот, сглаженное графическое изображение эмпирической функции распределения.

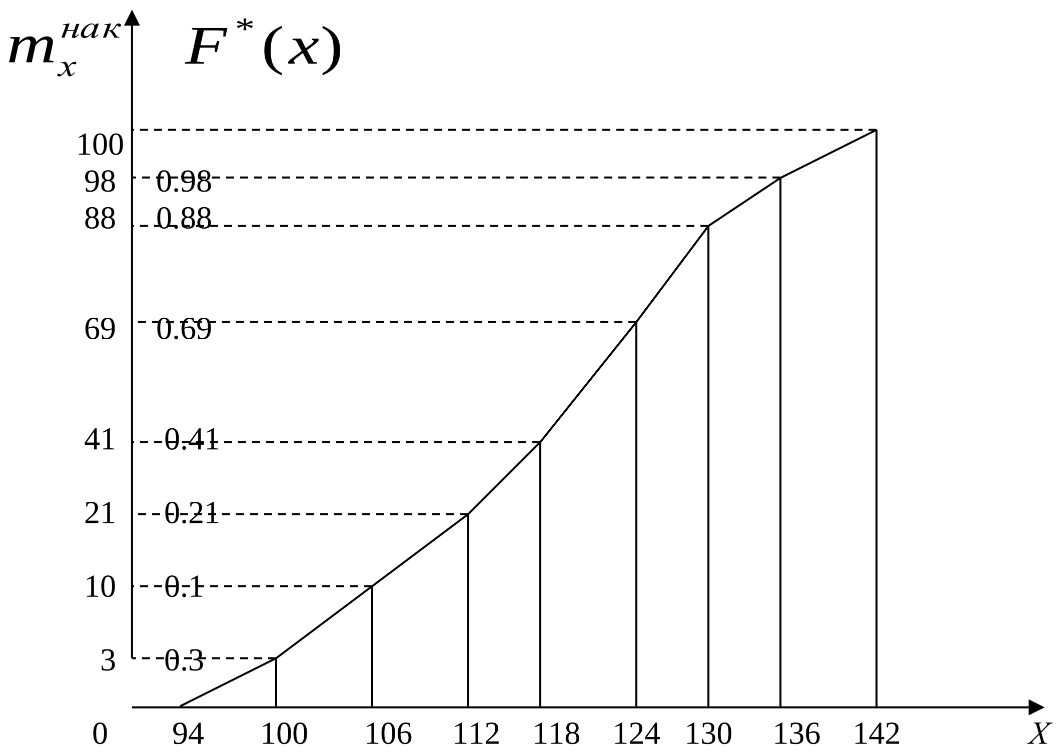


Рис. 1.5

При построении кумуляты в точке, соответствующей принимаемому значению, для дискретного ряда и в правом конце интервала для интервального ряда строится перпендикуляр, высота которого пропорциональна накопленной частоте, а затем верхние концы перпендикуляров соединяются между собой с помощью отрезков прямых.

Пример 7. Кумулята распределения рабочих по выработке (табл. 1.6) приведена на рисунке 1.5.

1.2. Примеры решения типовых задач

Задача 1.1. Задано следующее распределение 50 рабочих электротехнического цеха по тарифному разряду:

Тарифный разряд x_i	1	2	3	4	5	6	Σ
Частота (количество рабочих) m_i	2	3	6	8	22	9	50

Составить распределение относительных частот.

Решение.

Найдем относительные частоты по формуле (1.1):

$$p_1^* = \frac{m_1}{n} = \frac{2}{50} = 0{,}04\,; \qquad p_2^* = \frac{m_2}{n} = \frac{3}{50} = 0{,}06\,;$$

$$p_3^* = \frac{m_3}{n} = \frac{6}{50} = 0{,}12\,; \qquad p_4^* = \frac{m_4}{n} = \frac{8}{50} = 0{,}16\,;$$

$$p_5^* = \frac{m_5}{n} = \frac{22}{50} = 0{,}44\,; \qquad p_6^* = \frac{m_6}{n} = \frac{9}{50} = 0{,}18\,.$$

Контроль:

$$\sum_{i=1}^{6} p_i = 0{,}04 + 0{,}06 + 0{,}12 + 0{,}16 + 0{,}44 + 0{,}18 = 1.$$

Тогда искомое распределение относительных частот:

Тарифный разряд x_i	1	2	3	4	5	6
Относительная частота p_i^*	0,04	0,06	0,12	0,16	0,44	0,18

Задача 1.2. По данному распределению выборки требуется: 1) построить полигон частот; 2) построить полигон относительных частот; 3) найти и построить эмпирическую функцию распределения.

Значения x_i	1	4	5	8
Частоты m_i	10	5	20	15

Решение.

1) Для построения полигона частот заданного распределения отложим на оси абсцисс варианты x_i, а на оси ординат соответствующие им частоты m_i. Соединив точки (x_i ; m_i) отрезками прямых, получим искомый полигон частот (рис. 1.6).

? Запишите недостающие координаты вершин данного полигона частот: (1;10), (4;); (? , 20); ? ;). ? ?

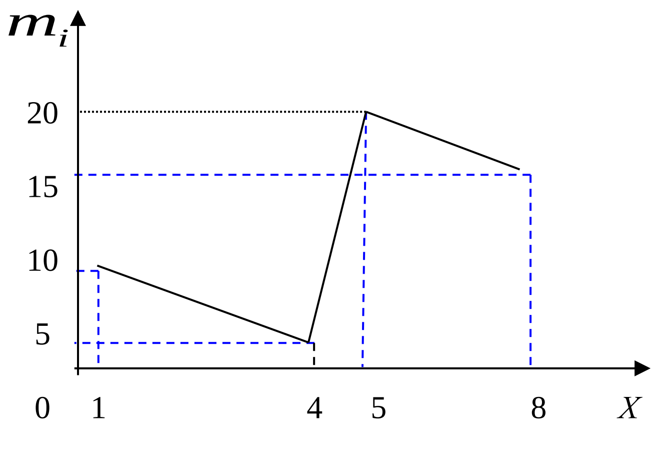


Рис. 1.6

2) Прежде чем построить полигон относительных частот, найдем распределение относительных частот. Объем заданной выборки:

$$n = \sum_{i=1}^{4} m_i = m_1 + m_2 + m_3 + m_4 = 10 + 5 + 20 + 15 = 50.$$

Относительные частоты найдём по формуле ? : $\boxed{?}$

$$p_1 = \frac{10}{50} = 0{,}2\,; \quad p_2 = \frac{?}{50} = 0{,}1; \quad p_3 = \frac{20}{?} = 0{,}4\,; \quad p_4 = \frac{?}{?} = 0{,}3.$$

Контроль: $\sum_{i=1}^{4} p_i = ? + ? + ? + ? = 1.$

Итак, получаем следующее распределение относительных частот:

x_i	1	4	5	8
m_i	0,2	0,1	0,4	0,3

Полигон относительных частот строится аналогично полигону частот, но на оси ординат откладываются относительные частоты $\tilde{p}_i$ (а не частоты m_i) (рис. 1.7).

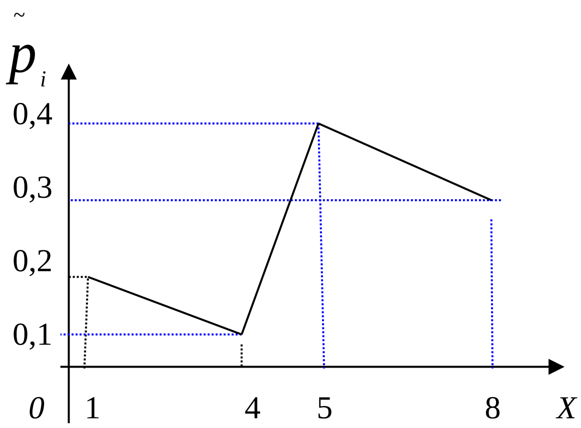


Рис. 1.7

$\boxed{?}$ Сравните графики на рис. 1.6 и 1.7. Сделайте вывод.

3) Найдём эмпирическую функцию распределения.

а) Наименьшая варианта равна единице, следовательно, $F^*(x)=0$ при $x \le 1$.

б) Значение $x<4$, а именно $x_1=1$, наблюдалось 10 раз, значит $F^*(x)=\frac{10}{50}=0{,}2$ при $1<x\le 4$.

в) Значения $x<5$, а именно $x_1=1$ и $x_2=4$, наблюдались 10+5=15 раз, поэтому $F^*(x)=\frac{15}{50}=0{,}3$ при $4<x\le 5$.

г) Значения $x<8$, а именно $x_1=1$, $x_2=4$ и $x_3=5$ наблюдались 10+5+20=35 раз, поэтому $F^*(x)=\frac{35}{50}=0{,}7$ при $5<X\le 8$.

д) Так как x =8 – наибольшая варианта, то $F^*(x)=1$ при x>8.

Итак, аналитическое выражение эмпирической функции распределения:

$$F^*(x)=\begin{cases} 0 & при \quad x\le 1; \\ 0{,}2 & при \quad x\le 4; \\ 0{,}3 & при \quad 4<x\le 5; \\ 0{,}7 & при \quad 5<x\le 8; \\ 1 & при \quad x>8. \end{cases}$$

График этой функции изображен на рис. 1.8.

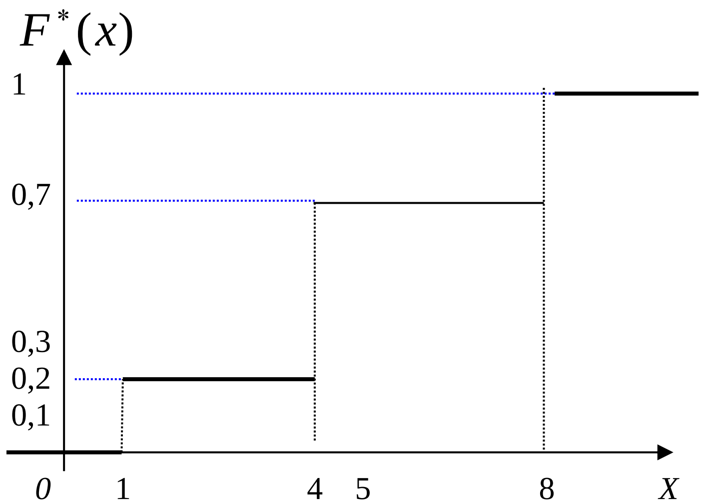


Рис. 1.8

$\boxed{?}$ Сравните процессы построения эмпирической функции распределения $F^*(x)$ для дискретного вариационного ряда и функции распределения *F(x)* для дискретной случайной величины.

Задача 1.3. По данному распределению выборки объёма *n*=100:

Интервал $x_i < x \le x_{i+1}$	1-11	11-21	21-31	31-41	41-51
Частоты m_i	10	20	8	50	12

Требуется построить: 1) гистограмму частот; 2) гистограмму относительных частот; 3) кумулятивную кривую.

Решение. 1)Для построения гистограммы частот построим на оси абсцисс, заданные интервалы длины *h* =10 (– как определ $\boxed{?}$ длину *h*?). Проведём над этими интервалами отрезки, параллельные оси абсцисс и находящиеся от неё на расстояниях, равных соответствующим плотностям частоты $\frac{m_i}{h}$ (рис.1.9). Соответствующие подсчёты для удобства проведём в таблице 1.8 (четвертый столбец).

Таблица 1.8

№	Интервалы $x_i < x \le x_{i+1}$	m_i	$\frac{m_i}{h}$	$\tilde{p}_i$	$\frac{\tilde{p}_i}{h}$	$F^*(x)$
1	**2**	**3**	**4**	**5**	**6**	**7**
1	1–11	10	$\frac{10}{10}=1$	$\frac{10}{100}=0,1$	$\frac{0,1}{10}=0,01$	0,1
2	11–21	20	$\frac{20}{10}=2$	$\frac{20}{?}=0,2$	$\frac{0,2}{10}=0,02$	0,1+0,2=0,3
3	21–31	8	$\frac{8}{10}=0,8$	$\frac{?}{100}=0,08$	$\frac{?}{10}=0,008$	0,3+?=0,38
4	31–41	50	$\frac{50}{10}=5$	$\frac{?}{?}=0,5$	$\frac{0,5}{?}=0,05$	?+?=0,88
5	41–51	12	$\frac{12}{10}=1,2$	$\frac{?}{?}=0,12$	$\frac{?}{?}=0,012$	?+?=1,00
Σ	–	100	–	$\boxed{?}$	–	–

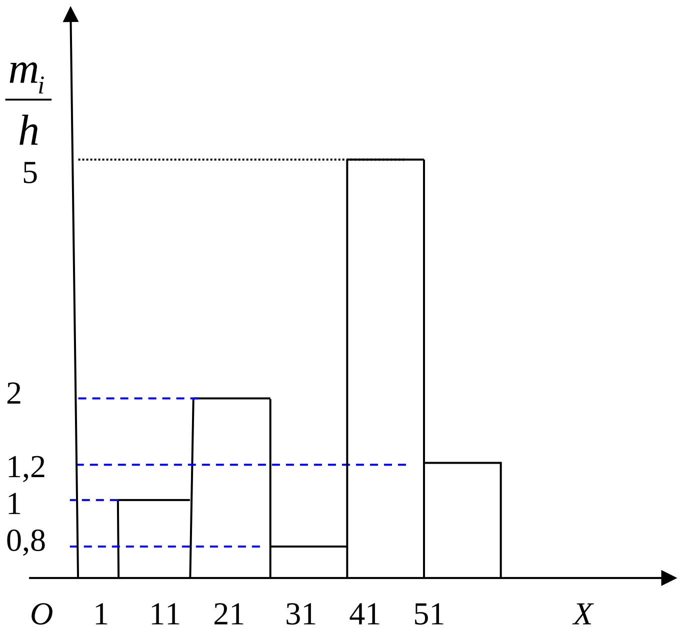


Рис. 1.9

2) Гистограмма относительных частот строится аналогично гистограмме частот (рис. 1.9), но отрезки, параллельные оси абсцисс, проводят на расстояниях, равным соответствующим плотностям относительной частоты $\frac{\tilde{p}_i}{h}$ (шестой столбец таблицы 1.8).

Например, над интервалом 1–11 проведём отрезок, параллельный оси абсцисс и находящийся от неё на расстоянии, равном $\frac{\tilde{p}_1}{h} = 0{,}01.$

Гистограмму относительных частот построить самостоятельно.

? 3) Значениями эмпирической функции распределения $F^*(x)$ являются накопленные частости (седьмой столбец табл. 1.8).

1.3. Задачи для самостоятельного решения

1.Выборка задана в виде распределения частот:

x_i	4	7	8	12
m_i	5	2	3	10

Найти распределение относительных частот.

2. Наблюдается число выигрышей в мгновенной лотерее. В результате наблюдения получены следующие значения выигрышей (тыс. руб.):

0,1, 0, 0, 5, 0, 10, 0, 1, 0, 0, 1, 5, 1, 0, 0, 0, 1, 0, 1, 0, 0, 0, 0, 5, 0, 5, 0, 0, 1, 1, 1, 5, 10, 0, 1, 1, 0, 5, 0, 0, 0, 0, 1, 0, 1, 0, 5, 0, 0, 0, 0, 1, 0.

Составить вариационный ряд случайной величины X – выигрыша в мгновенной лотерее.

3. Имеются следующие данные (среднегодовая численность работающих) по отдельным предприятиям за отчетный период:

655, 815, 925, 810, 820, 795, 840, 445, 932, 1200, 1140, 1290, 1050, 1225, 1140, 1521, 1460, 1600, 955, 562, 485, 515, 622, 400, 750, 1650.

Построить по этим данным интервальный вариационный ряд с равными интервалами, взяв за первый интервал 400 – 500, а за второй 500 – 600 и т.д.

4. Найти эмпирическую функцию распределения по следующему вариационному ряду:

x_i	1	3	7	9	12
m_i	2	10	4	24	10

5. Найти эмпирическую функцию распределения по следующему интервальному вариационному ряду:

i	1	2	3	4	5	6	7	8	9
$x_i < X \le x_{i+1}$	0-2	2-4	4-6	6-8	8-10	10-12	12-14	14-16	16-18
m_i	6	4	2	18	29	11	10	17	3

6. Дано распределение случайной величины X – числа сделок на фондовой бирже за квартал; n = 400 (инвесторов):

x_i	0	1	2	3	4	5	6	7	8	9	10
m_i	146	97	73	34	23	10	6	3	4	2	2

Требуется построить:

1) полигон; 2) кумуляту; 3)эмпирическую функцию распределения.

7. Дано распределение признака X – удоя коров на молочной ферме за лактационный период (в ц); n = 100 (коров):

x_i	4-6	6-8	8-10	10-12	12-14	14-16	16-18	18-20	20-22	22-24	24-26
m_i	1	3	6	11	15	20	14	12	10	6	2

Требуется построить:

1)гистограмму; 2) кумуляту; 3) эмпирическую функцию распределения.

---

8. Выборка задана в виде распределения частот:

x_i	2	5	7	8
m_i	1	3	2	4

Требуется: 1) найти распределение относительных частот;

2) записать эмпирическую функцию распределения и построить ее график; 3) построить полигоны частот и относительных частот.

9. В супермаркете проводились наблюдения над числом X покупателей, обратившихся в кассу за один час. Наблюдения в течение 30 часов дали следующие результаты:

70, 75, 100, 120, 75, 60, 100, 120, 70, 60, 65, 100, 65, 100, 70, 75, 60, 100, 100, 120, 70, 75, 70, 120, 65, 70, 75, 70, 100, 100.

Составить ряд распределения частот и относительных частот.

10. Продукция (в млн руб.) за отчетный период у 25 сельхозпредприятий составила:

4,2; 6,5; 7,3; 10,5; 6,0; 9,2; 6,8; 3,3; 10,2; 11,5; 11,9; 14,1; 11,2; 14,2; 17,7; 17,6; 18,5; 19,2; 7,3; 3,2; 4,1; 3,7; 4,0; 5,1; 6,5.

Построить интервальный вариационный ряд, взяв за интервалы 13,2 – 14,2; 14,2 – 15,2 и т.д. Начертить гистограмму.

11. Дано распределение признака X – недельного дохода жителя Приморского края (в руб.); n = 1000 (человек):

x_i	Менее 500	500-1000	1000-1500	1500-2000	2000-2500	Более 2500
m_i	58	96	239	328	147	132

Требуется построить:

1) гистограмму;

2) кумуляту;

3) эмпирическую функцию распределения.

12. Построить дискретный вариационный ряд и начертить полигон распределения 60 абитуриентов по числу баллов, полученных ими на вступительных экзаменах:

20, 19, 22, 24, 21, 18, 23, 17, 20, 16, 15, 23, 21, 24, 21, 18, 23, 21, 19, 20, 24, 21, 20, 18, 17, 22, 20, 16, 22, 18, 20, 17, 21, 17, 19, 20, 20, 21, 18, 22, 23, 21, 25, 22, 20, 19, 21, 24, 23, 21, 19, 22, 21, 19, 20, 23, 22, 25, 21, 21.

13. Построить дискретный вариационный ряд и начертить полигон распределения 45 пар мужской обуви, проданных магазином за день:

39 , 41, 40, 42, 41, 40, 42, 44, 40, 43, 42, 41, 43, 39, 42, 41, 42, 39, 41, 37, 43, 41, 38, 43, 42, 41, 40, 41, 38, 44, 40, 39, 41, 40, 42, 40, 41, 42, 40, 43, 38, 39, 41, 41, 42.

14. Имеются следующие данные о распределении предприятий края по величине роста выработки на одного рабочего:

Выработка на одного рабочего в отчетном году (в % к предыдущему году)	80- 90	90-100	100-110	110-120	120-130	130-140	140 и более
Число предприятий	8	15	46	29	13	3	3

Приняв ширину первого и последнего интервалов равной 10 %, построить:

1) кумулятивную кривую;

2) гистограмму.

15. Результаты выборочного обследования среднемесячного дохода 100 человек, работающих в бюджетной сфере, приведены в следующей таблице:

Среднемесячный доход (руб.)	Число человек
до 3000	14
от 3000 до 4000	21
от 4000 до 6000	23
от 6000 до 8000	18
от 8000 до 10000	10
от 10000 до 15000	9
более 15000	5

Построить гистограмму, кумуляту, эмпирическую функцию распределения.

16. По выборке С (*Таблица* 1 *Приложения* 1) требуется:

1.составить дискретный вариационный ряд;

2. вычислить относительные частоты и накопленные частости;

3. построить полигон распределения;

4. составить эмпирическую функцию распределения;

5. построить график эмпирической функции распределения.

17. По выборке Д (*Таблица2 Приложения* 1) требуется:

1. составить интервальный вариационный ряд;

2. построить гистограмму;

3. построить кумулятивную кривую.

18. Из генеральной совокупности сельхозпредприятий Приморского края сделана случайная выборка из 80 хозяйств (*Таблица* 3 *Приложения* 1). Исследуемый признак *X*– количество реализованного мяса свиней (в центнерах). Требуется:

1. составить вариационный ряд;

2.построить графики вариационного ряда;

3. построить график эмпирической функции распределения.

19. Из генеральной совокупности сельхозпредприятий Приморского края, занимающихся производством зерновых, взята выборка из 70 хозяйств (*Таблица* 4 *Приложения* 1) необходимо:

1. составить вариационный ряд; 2. построить гистограмму и кумуляту.

20. По выборке из 50 сельхозпредприятий Приморского края, выращивающих зерновые (*Таблица 5 Приложения* 1), требуется:

1. составить вариационный ряд;
2. вычислить относительные частоты и накопленные частости;
3. построить графики вариационного ряда.

Тест 1

1. Генеральной совокупностью называется:

а) совокупность всех возможных однородных предметов или явлений, над которыми проводятся наблюдения;

б) совокупность предметов или явлений, отобранная из выборочной совокупности;

в) совокупность генералов;

г) совокупность методов получения, описания и обработки опытных данных.

2. Выборка из генеральной совокупности задана в виде распределения частот. Заполните в таблице недостающие характеристики случайной величины:

Значения x_i	3	7	11	15	19
Частоты m_i	2	7	11	8	2
Относительные частоты $\tilde{p}_i$					
Накопленные частоты $m_x^{нак}$					
Накопленные частости $p_x^{нак}$					

3. Значение случайной величины x_i, соответствующее отдельной группе сгруппированного ряда наблюдаемых данных, называется:

а) частотой;

б) относительной частотой;

в) вариантой;

г) кумулятой.

4. Дана выборка оценок 15 студентов: 3,3,4,4,5,3,3,3,4,5,4,4,5,3,3.

Тогда дискретный вариационный ряд имеет вид:

а)

x_i	3	4	5
m_i	7	5	3

б)

x_i	3	4	5
m_i	3	5	7

в)

x_i	3	5	7
m_i	5	3	4

г)

x_i	1	2	3
m_i	7	5	3

5.Относительной частотой варианты x_i называется:

а) произведение m_i числа повторения на x_i объем выборки n:

$$\tilde{p}_i = m_i \cdot n;$$

б) сумма m_i числа повторения x_i и объема выборки n:

$$\tilde{p}_i = m_i + n;$$

в) отношение объема выборки n к числу повторения m_i варианты x_i:

$$\tilde{p}_i = \frac{n}{m_i};$$

г) отношение m_i числа повторения x_i к объему выборки n:

$$\tilde{p}_i = \frac{m_i}{n}.$$

6. Для распределения выборки:

x_i	1	4	6
m_i	10	15	25

Эмпирическая функция имеет вид:

$$F^*(x) = \begin{cases} 0 & при \quad x \le 1; \\ a & при \quad 1 < x \le 4; \\ b & при \quad 4 < x \le 6; \\ 1 & при \quad x > 6. \end{cases}$$

Тогда значение выражения $(5a + 2b) = \ldots$

7. Ранжированная совокупность вариант с соответствующими им частотами или относительными частотами называется:

а) статистической совокупностью;

б) выборочной совокупностью;

в) дискретным вариационным рядом;

г) интервальным вариационным рядом.

8. Для выборки объема $n = 100$ оптимальное число интервалов (по формуле Стерджеса) равно:

а) 5; б) 6; в) 8; г) 10.

9. Эмпирической функцией распределения называется функция $F^*(x)$, определяющая для каждого значения x относительную частоту события:

а) $X < x, \quad т.е. \quad F^*(x) = \tilde{P}(X < x)$;

б) $X > x, \quad т.е. \quad F^*(x) = \tilde{P}(X > x)$;

в) $X = \sum x_i, \qquad т.е. \quad F^*(x) = \tilde{P}\left(X = \sum x_i\right);$

г) $X \geq x, \qquad т.е. \quad F^*(x) = \tilde{P}(X \geq x).$

10. Накопленной частотой $m_x^{нак}$ в точке x называется:

а) суммарная частота членов статистической совокупности со значениями признака, меньшими, чем x;

б) произведение членов статистической совокупности со значениями признака, меньшими, чем x;

в) отношение суммарной частоты членов совокупности со значениями признака, меньшими, чем x, к объему выборки n;

г) наибольшая частота членов статистической совокупности.

11. Установить соответствие:

1. Полигон частот – это	а) ломаная, отрезки которой соединяют точки $(x_i;\ \tilde{p}_i)$ $\left(i = \overline{1,k}\right)$;
2.Полигон относительных частот – это	б) ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению $\tilde{p}_i / h$;
3. Гистограмма частот – это	в) ломаная, отрезки которых соединяют точки $(x_i;\ p_i)$ $\left(i = \overline{1,k}\right)$;
4.Гистограмма относительных частот – это	г) ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению m_i / h.

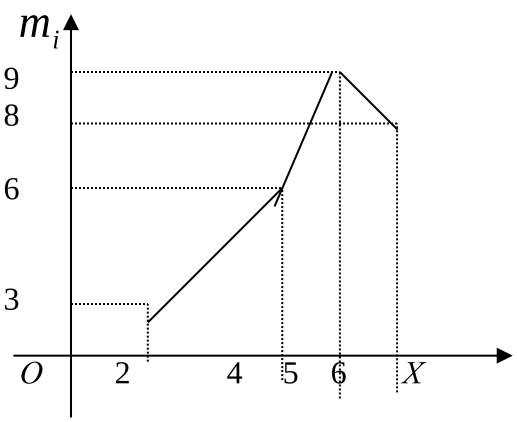
12. Построить гистограмму частот следующего интервального вариационного ряда:

Интервалы	3 – 6	6 – 9	9 – 12	12 –15
Частоты m_i	9	10	25	6

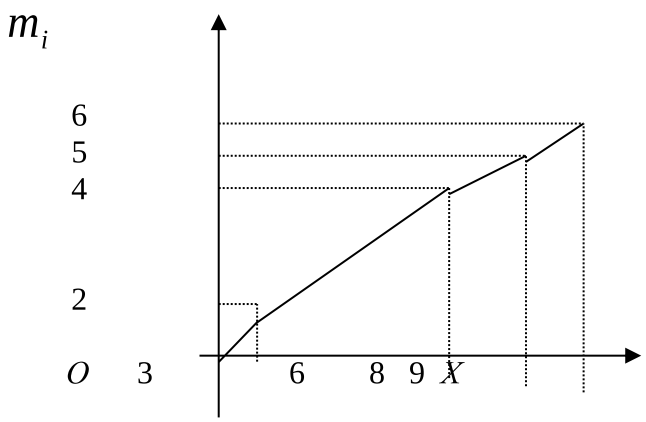
13. Полигон частот выборки

x_i	3	6	8	9
m_i	2	4	6	5

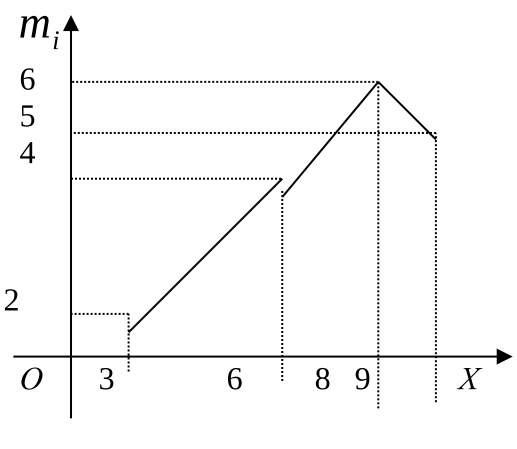
представлен на рисунке:



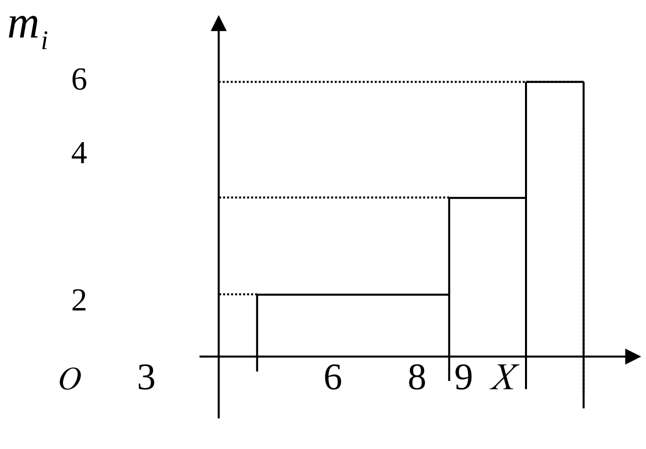
а)



б)



в)



г)

14. Минимальное значение выборки $x_{\min} = 25$, длина интервала $h = 4$

Тогда за начало первого интервала целесообразно брать $x_{нач} = \ldots$

15. Математическая статистика – это:

а) раздел экономической теории;

б) наука, занимающаяся разработкой методов получения описания и обработки опытных данных с целью изучения закономерностей случайных массовых явлений;

в) наука, изучающая поведение экономических систем;

г) наука, занимающаяся разработкой методов получения описания обработки эмпирических данных с целью получения закономерностей отдельных событий.

Вопросы для самопроверки по главе 1

1. Понятие генеральной и выборочной совокупности.

2. Повторная и бесповторная выборки. Репрезентативная выборка.

3. Понятие варианты и частоты (веса) варианты.

4. Относительная частота варианты.

5. Построение таблицы статистического распределения.

6. Дискретный и интервальный вариационные ряды: понятие и алгоритмы построения.

8. Накопленная частота и накопленная относительная частота.

9. Эмпирическая функция распределения.

10. Полигон и гистограмма (частот и относительных частот).

12. Кумулятивная кривая.

ГЛАВА 2. ЧИСЛОВЫЕ ХАРАКТЕРИСТИКИ ВАРИАЦИОННОГО РЯДА (ВЫБОРКИ). ЭЛЕМЕНТЫ ТЕОРИИ ОЦЕНОК

2.1.Необходимый теоретический минимум

На основе составленного для статистической совокупности ряда распределения можно установить некоторые характеристики этой совокупности, выявляющие как особенности расположения и концентрации значений варьирующего признака (среднее арифметическое, медиана, мода), так и характеризующие меру варьирования признака, степени его колеблемости (размах вариации, среднее линейное отклонение, дисперсия, среднее квадратическое отклонение).

I. Характеристики уровня вариационного рядов

1. Выборочное среднее (среднее арифметическое):

а) ***выборочное среднее (среднее арифметическое) простое***:

$$\overline{x}_в = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{1}{n}\left(x_1 + x_2 + x_3 + \ldots + x_n\right), \tag{2.1}$$

где x_i $\left(i = \overline{1,n}\right)$ - варианты признака (элементы выборки),

n – объем выборки.

- Если частоты вариантов признака различны (т.е. имеют различный удельный вес во всем объеме совокупности), то вычисляется среднее арифметическое, взвешенное по частотам.

б) выборочное среднее (среднее арифметическое) взвешенное:

$$\overline{x}_в = \frac{1}{n}\cdot\sum_{i=1}^{k} x_i \cdot m_i = \frac{1}{n}\left(x_1 m_1 + x_2 m_2 + \ldots + x_k m_k\right), \tag{2.2}$$

где x_i - варианты случайной величины (признака),

m_i $\left(i = \overline{1,k}\right)$ – соответствующие частоты,

k – количество вариантов, n – объем выборки.

- Если ряд интервальный, то при вычислении среднего арифметического в роли x_i в формулах (2.1) и (2.2) представителем каждого интервала берется его середина.
- Для упрощения счета взвешенного выборочного среднего используется следующая формула:

$$\bar{x}_в = \frac{h}{n} \cdot \sum_{i=1}^{k} \left(\frac{x_i - c}{h} \right) \cdot m_i + c, \qquad \textbf{(2.3)}$$

где x_i, m_i, k, n – имеют тот же смысл, что и в формуле (2.2), h – шаг таблицы (т. е. интервал между соседними вариантами); c - произвольное число (близкое к ожидаемому среднему, или варианта с максимальной частотой).

- Если среднее арифметическое находится для генеральной совокупности, то его называют генеральной средней и обозначают $a = \bar{x}_г$.

Пример. Найти выборочное среднее распределения, заданного интервальным рядом (табл. 2.1).

Таблица 2.1

$x_i < x \le x_{i+1}$	10-20	20-30	30-40	40-50	50-60
m_i	12	18	27	25	18

Решение. 1) При расчете обычным способом (т.е. по формуле (2.2)) получаем:

$$\bar{x}_в = \frac{15 \cdot 12 + 25 \cdot 18 + 35 \cdot 27 + 45 \cdot 25 + 55 \cdot 18}{12 + 18 + 27 + 25 + 18} = 36{,}9.$$

2) Покажем теперь методику расчета $\bar{x}_в$ по формуле (2.3). Возьмем в качестве $C = 35$ (т.е. середину интервала, которому соответствует

максимальная частота); $d = 10$ (длина интервала). Соответствующие расчеты проведем в таблице 2.2.

Таблица 2.2.

№	Интервалы $x_i < x \le x_{i+1}$	Частота m_i	Середина интервала x_i	$\frac{x_i - c}{h}$	$\left(\frac{x_i - c}{h}\right) \cdot m_i$
1	10-20	12	15	$\frac{15-35}{10} = -2$	$-2 \cdot 12 = -24$
2	20-30	18	25	$\frac{25-?}{10} = -1$	$-1 \cdot 18 = -18$
3	30-40	27	35=C	? $= 0$	$0 \cdot 27 = 0$
4	40-50	25	45	? $= 1$	$1 \cdot ? = 25$
5	50-60	18	55	? $= 2$	$? \cdot ? = 36$
	$\sum$	100	–	–	$19 = \sum_{i=1}^{5} \left(\frac{x_i - c}{h}\right) \cdot m_i$

Итак, из таблицы 2.2 видно, что: $\sum_{i=1}^{5} \left(\frac{x_i - c}{h}\right) \cdot m_i = 19$.

Тогда получаем: $\bar{x}_в = \frac{10}{100} \cdot 19 + 35 = 36{,}9.$

- Рассмотренный метод расчета выборочного среднего называется *методом условного нуля*. Выбор условного нуля (т.е. величины C) на результат не влияет.

$\boxed{?}$ Где Вы ранее (в курсе теории вероятностей) встречались с методом условного нуля?

? Математическое ожидание случайной величины и среднее арифметическое выборки как можно соотнести друг с другом?

2. Мода

а) Для дискретного вариационного ряда ***модой*** M_0 выборки является значение, имеющее максимальную частоту.

б) Для интервального вариационного ряда ***мода*** M_0 вычисляется по приближенной формуле:

$$M_0 \approx x_0 + h \cdot \frac{m_i - m_{i-1}}{(m_i - m_{i-1}) + (m_i - m_{i+1})}, \qquad \textbf{(2.4)}$$

где:

x_0 – начало модального интервала, т.е. интервала, имеющего максимальную частоту, h – длина модального интервала, m_i - частота модального интервала, m_{i-1} и m_{i+1} – частоты соответственно предшествующего и последующего за модальным интервалов.

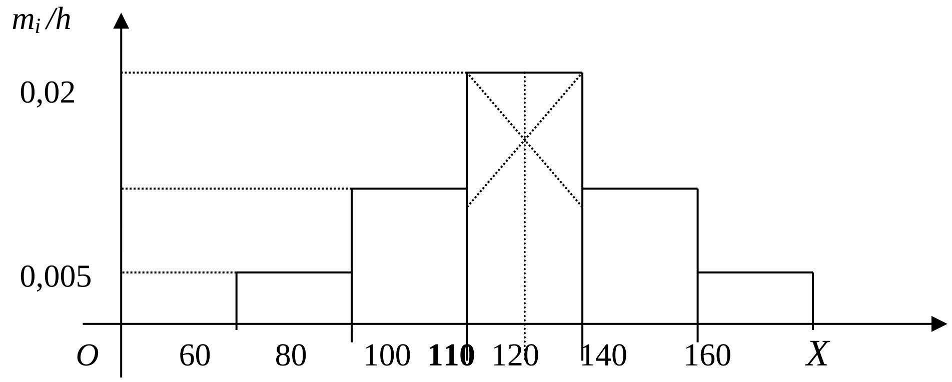


Рис. 2.1

• Моду также можно определить графически, используя соответствующую гистограмму (частот или относительных частот). Действительно, на гистограмме интервального ряда легко находится интервал, у которого частота максимальна. Мода находится внутри его (на пересечении отрезков, проведенных из вершин прямоугольника, соответствующего модальному интервалу, до вершин «предмодального» и «послемодального» прямоугольников). Например, для интервального ряда, гистограмма которого представлена на рис.2.1, мода равна $M_0 = 110$.

3. Медиана

Медиана выборки – это значение серединного элемента вариационного ряда.

а) Для дискретного вариационного ряда ***медиана*** M_e находится по формуле:

$$M_e = \begin{cases} \frac{1}{2} \cdot \left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1} \right), & \text{если } n - \text{четное}; \\ x_{\frac{n+1}{2}}, & \text{если } n - \text{нечетное}. \end{cases} \quad \textbf{(2.5)}$$

б) Для интервального вариационного ряда ***медиана*** M_e вычисляется по формуле:

$$M_e = x_0 + h \cdot \frac{\frac{n}{2} - T_{i-1}}{m_i}, \quad \textbf{(2.6)}$$

где: x_0 - начало медианного интервала, (т.е. интервала, в котором находится серединный элемент); h – длина медианного интервала; n – объем выборки; T_{i-1} – сумма частот интервалов, предшествующих медианному; m_i - частота медианного интервала.

- С геометрической точки зрения медиана – это точка, в которой площадь гистограммы делится пополам. Поэтому медиану можно определить и графически, используя кумулятивную кривую. Например, для интервального ряда, кумулята которого представлена на рис.2.2, медиана равна $M_e = 150$.

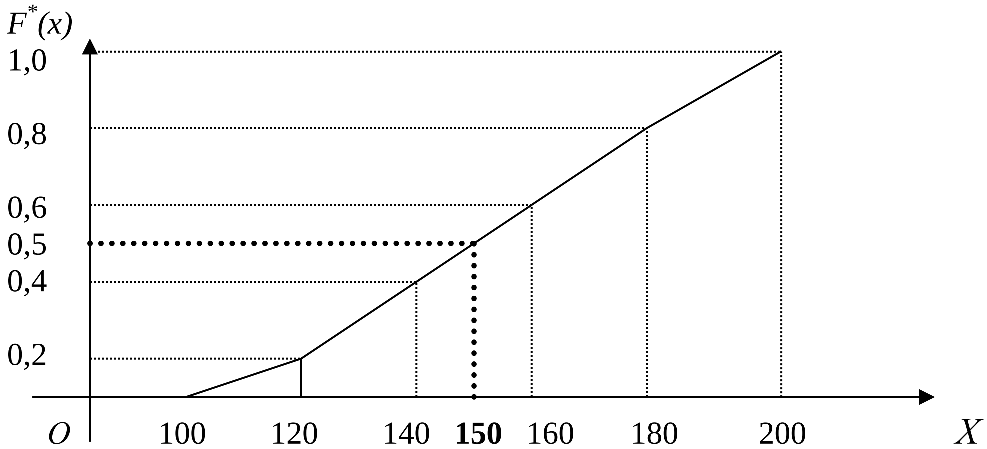


Рис. 2.2

II. Показатели колеблемости вариационных рядов

1. ***Размах вариации:*** $$R = x_{\max} - x_{\min}\ ,\qquad \textbf{(2.7)}$$

где $x_{\max}$ – максимальное, $x_{\min}$ – минимальное значения выборки.

2. ***Среднее линейное отклонение*** – это среднее арифметическое, исчисленное из абсолютных отклонений отдельных вариант (x_i) от их выборочной средней ($\bar{x}_в$). Различают:

а) *среднее линейное отклонение (простое):*

$$d = \frac{1}{n}\cdot\sum_{i=1}^{k}\left|x_i - \bar{x}_в\right|;\qquad \textbf{(2.8)}$$

б) *среднее линейное отклонение (взвешенное):*

$$d = \frac{1}{n}\cdot\sum_{i=1}^{k}\left|x_i - \bar{x}\right|\cdot m_i.\qquad \textbf{(2.9)}$$

3. ***Выборочная дисперсия*** $D_в$ – это среднее арифметическое квадратов отклонений отдельных вариант (x_i) от их выборочной средней ($\bar{x}_в$). Бывает:

а) *выборочная дисперсия (простая):*

$$D_в = \frac{1}{n}\cdot\sum_{i=1}^{k}\left(x_i - \bar{x}_в\right)^2\qquad \textbf{(2.10)}$$

или

$$D_в = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i^2 - \bar{x}_в^{\,2}; \qquad \textbf{(2.11)}$$

б) *выборочная взвешенная дисперсия*:

$$D_в = \frac{1}{n} \cdot \sum_{i=1}^{k} \left(x_i - \bar{x}_в\right)^2 \cdot m_i \qquad \textbf{(2.12)}$$

или

$$D_в = \frac{1}{n} \cdot \sum_{i=1}^{k} x_i^2 \cdot m_i - \bar{x}_в^{\,2}. \qquad \textbf{(2.13)}$$

? Поясните смысл обозначений k, n, m_i в формулах (2.8) – (2.13).

- Если дисперсия находится для генеральной совокупности, то её называют *дисперсией генеральной совокупности*, обозначают D, и находят по формулам, аналогичным формулам (2.10) – (2.13).

в) *исправленная дисперсия*: $$S^2 = \frac{n}{n-1} \cdot D_в. \qquad \textbf{(2.14)}$$

4. ***Выборочное среднее квадратическое отклонение*** $\sigma_в$ – квадратный корень из дисперсии. Различают:

а) *выборочное среднее квадратическое отклонение простое*:

$$\sigma_в = \sqrt{\frac{1}{n} \sum_{i=1}^{k} \left(x_i - \bar{x}_в\right)^2} \qquad \textbf{(2.15)}$$

или

$$\sigma_в = \sqrt{\frac{1}{n} \sum_{i=1}^{k} x_i^2 - \bar{x}_в^{\,2};} \qquad \textbf{(2.16)}$$

б) *выборочное среднее квадратическое отклонение взвешенное*:

$$\sigma_в = \sqrt{\frac{1}{n}\sum_{i=1}^{k}\left(x_i - \bar{x}_в\right)^2 \cdot m_i} \qquad \textbf{(2.17)}$$

или

$$\sigma_в = \sqrt{\frac{1}{n}\sum_{i=1}^{k} x_i^2 \cdot m_i - \bar{x}_в^{\ 2}}. \qquad \textbf{(2.18)}$$

? Поясните смысл обозначений $\bar{x}_в$, k, n, m_i в формулах (2.15) – (2.18).

5. ***Выборочный коэффициент вариации*** – это отклонение выборочного среднего квадратического отклонения к выборочному среднему, выраженное в процентах:

$$V_в = \frac{\sigma_в}{\bar{x}_в} \cdot 100(\%) \qquad \textbf{(2.19)}$$

- Пользуясь вариационными рядами и полученными на их основе средними показателями (выборочной средней и выборочным средним квадратическим отклонением) можно давать сравнительную оценку различным процессам. Например, с их помощью можно охарактеризовать качество работы предприятий в зависимости от некоторых условий, сравнить эффективность различных производственных технологий и т.п.

II. Точечные оценки и их свойства (виды)

- Генеральные совокупности характеризуются некоторыми постоянными числовыми характеристиками распределения. По выборкам можно найти оценки этих характеристик. Из-за случайности выборок значения оценок одной числовой характеристики, вычисленные по разным выборкам из одной и той же генеральной совокупности, бывают, как правило, различными.

Пусть θ – неизвестный параметр распределения (числовой характеристики генеральной совокупности X), а θ_n – его оценка.

Определение 2.1. ***Статистической оценкой*** θ_n неизвестного параметра θ теоретического распределения называют функцию $f(x_1,\ x_2,\ldots x_n)$ от наблюдаемых случайных величин $x_1, x_2, \ldots, x_n$.

• Оценки неизвестного параметра можно находить различными способами. И для того, чтобы установить, какая из оценок лучше, надо знать основные свойства (виды) оценок.

Определение 2.2. Оценка θ_n неизвестного параметра θ, определяющая одну точку на числовой оси, называется *точечной оценкой*.

Виды точечных оценок:

– ***несмещенная*** – это такая оценка θ_n, среднее значение (математическое ожидание) которой равно оцениваемому параметру θ при любом объеме выборки, т.е. $M(\theta_n) = \theta$; **(2.20)**

– если условие (2.20) не выполняется, то оценку называют ***смещенной***, при этом смещение определяется как разность $\left(M(\theta_n) - \theta\right)$.

• Смещенная оценка θ_n (полученная по разным выборкам) будет в среднем либо завышать, либо занижать значения θ.

– ***эффективная*** – это такая оценка θ_n, которая при заданном объеме выборки имеет наименьшую дисперсию;

– ***состоятельная*** – это такая оценка θ_n, которая при увеличении объема выборки, сходится по вероятности к истинному значению параметра.

- Для состоятельных оценок значительные ошибки при оценивании маловероятны.
- Из неравенства Чебышева непосредственно следует, что если дисперсия несмещенной оценки при $n \to \infty$, стремится к нулю, то такая оценка также является и состоятельной.

$\boxed{?}$ Закончите запись неравенство Чебышева применительно к теории оценок: $P(|\theta_n - \theta| < \xi) \geq 1 - \ldots$.

Некоторые оценки параметров генеральной совокупности:

1. Выборочная средняя $(\bar{x}_в)$ – это *несмещенная* оценка генеральной средней (математического ожидания).

2. Выборочная дисперсия при малом объеме выборки $(D_в)$– это *смещенная* оценка генеральной дисперсии.

3. Исправленная выборочная дисперсия S^2 – это *несмещенная* оценка генеральной дисперсии.

2.2.Примеры решения типовых задач

Задача 2.1. Из генеральной совокупности извлечена выборка объема $n = 60$:

x_i	3	4	6	26
m_i	8	38	12	2

I. Требуется найти числовые характеристики выборки:

1. выборочную среднюю;
2. моду;
3. медиану;

4. выборочную дисперсию;

II. Определить несмещенную оценку:

1. генеральной средней;

2. генеральной дисперсии.

Решение.

I. **1**. Выборочную среднюю найдем по формуле (2.2) для взвешенной средней (так как частоты вариантов признака различны). Все необходимые расчеты приведены в таблице 2.3.

Таблица 2.3.

i	x_i	m_i	$x_i \cdot m_i$	$x_i^2 \cdot m_i$	$m_x^{нак}$
1	**2**	**3**	**4**	**5**	**6**
1	3	8	$3 \cdot 8 = 24$	$3^2 \cdot 8 = 72$	8
2	4	38	$4 \cdot ? = 152$	$4^2 \cdot 38 = 608$	$8 + 38 = 46$
3	6	12	$? \cdot ? = 72$	?	$46 + 12 = 58$
4	26	2	$? \cdot ? = 52$	?	$58 + 2 = 60$
$\sum$	–	60	$300 \to \sum_{i=1}^{4} x_i m_i$	$2464 \to \sum_{i=1}^{4} x_i^2 \cdot m_i$	–

Итак, имеем:

$$\bar{x}_в = \bar{x} = \frac{1}{60} \cdot \sum_{i=1}^{4} x_i \cdot m_i = \frac{1}{60} \cdot 300 = 5.$$

(Требуемая сумма произведений $x_i \cdot m_i$ – есть сумма элементов четвертого столбца табл. 2.3). Значит, $\bar{x}_в = 5$.

2. Заданный вариационный ряд дискретный, для которого наибольшая частота $m_2 = 38$ отвечает значению признака $x_2 = 4$. Поэтому мода $M_0 = 4$.

3. Для нахождения медианы выборки строим ряд накопленных частот или кумулятивный ряд (шестой столбец таблицы 2.3). По формуле (2.5) при $n = 60$ получаем:

$$M_e = \frac{1}{2}\left(x_{\frac{n}{2}} + x_{\frac{n}{2}+1}\right) = \frac{1}{2}\left(x_{\frac{60}{2}} + x_{\frac{60}{2}+1}\right) = \frac{1}{2}\left(x_{30} + x_{31}\right).$$

Здесь, x_{30} и x_{31} – соответственно 30-ый и 31-ый члены вариационного ряда. Из кумулятивного ряда видно, что $x_{30} = x_{31} = 4$.

4. Выборочную дисперсию найдем по формуле (2.13) для дисперсии взвешенной: $D_в = \frac{1}{60}\sum_{i=1}^{4} x_i^2 \cdot m_i - \overline{x}_в^2.$

Здесь $\sum_{i=1}^{4} x_i^2 \cdot m_i = 2464$ (данная сумма найдена в пятом столбце табл. 2.3), а $\overline{x_в}^2 = 5^2 = 25.$ Итак, получаем:

$$D_в = \frac{1}{60} \cdot 2464 - 25 \approx 16{,}067.$$

II. **1.** Несмещенной оценкой генеральной средней является выборочная средняя, то есть $\overline{x}_в = 5.$

2.Несмещенной оценкой генеральной дисперсии является дисперсия исправленная, которую вычислим по формуле (2.14):

$$S^2 = \frac{n}{n-1} \cdot D_в = \frac{60}{60-1} \cdot 16{,}067 \approx 16{,}339.$$

Ответ. I. 1. $\overline{x}_в = 5$; 2. $M_0 = 4$; 3. $Me = 4$; 4. $D_в \approx 16{,}067$;

II. 1. $\overline{x}_в = 5$; 2. $S^2 = 16{,}339.$

Задача 2.2. На опытном поле испытываются два вида удобрений. Производится ряд опытов по определению влияния каждого из этих удобрений на урожайность. В результате были получены следующие ряды распределения урожайности:

I вид удобрений		II вид удобрений	
Урожайность (ц/га)	Площадь, га	Урожайность (ц/га)	Площадь, га
16-17	1	14-16	1
17-18	2	16-18	2
18-19	3	18-20	4
19-20	7	20-22	5
20-21	5	22-24	4
21-22	4	24-26	3
22-23	1	26-28	1
23-24	1		

Выяснить, какой вид удобрений в данных условиях более эффективен.

Решение. Чтобы определить, какой вид удобрений более эффективен, необходимо сравнить средние значения урожайности в данной серии опытов.

1. Найдем среднее значение урожайности при внесении I-го вида удобрений (табл. 2.4).

Таблица 2.4

i	Интервалы	Середина интервалов x_i	m_i	$x_i \cdot m_i$
1	16-17	16,5	1	$16{,}5 \cdot 1 = 16{,}5$
2	17-18	17,5	2	$17{,}5 \cdot 2 = ?$
3	18-19	18,5	3	$18{,}5 \cdot 3 = ?$
4	19-20	19,5	7	136,5
5	20-21	20,5	5	102,5
6	21-22	21,5	4	?
7	22-23	22,5	1	?
8	23-24	23,5	1	?
$\sum$	–	–	24	478

По формуле (2.2), получаем: $\bar{x}_I = \frac{1}{24} \cdot \sum_{i=1}^{8} x_i \cdot m_i = \frac{478}{24} \approx 19{,}92(ц/га)$.

2. Вычислим теперь среднюю урожайность при внесении II-го вида удобрений (табл. 2.5).

Таблица 2.5

i	Интервалы	Середина интервалов x_i	m_i	$x_i \cdot m_i$
1	14-16	15	1	$15 \cdot 1 = 15$
2	16-18	17	2	$17 \cdot 2 = ?$
3	18-20	19	4	$? \cdot 4 = 76$
4	20-22	21	5	$? \cdot 5 = ?$
5	22-24	23	4	?
6	24-26	25	3	?
7	26-28	27	1	27
$\sum$	-	-	20	?

Тогда получаем, что $\bar{x}_{II} = \frac{1}{20} \cdot \sum_{i=1}^{7} x_i \cdot m_i = \frac{1}{20} \cdot ? \approx 21{,}2(ц/га)$.

3. *Вывод*: так как $\bar{x}_{II} > \bar{x}_I$, то второй вид удобрений более эффективен.

Ответ. Второй вид удобрений более эффективен

Задача 2.3. Из генеральной совокупности сельхозпредприятий Приморского края сделана выборка из 100 хозяйств. Для исследуемого признака X – величина удоя за лактационный период на одну голову (в кг) – данные сведены в следующую таблицу:

Величина удоя, кг	476-1006	1006-1536	1536-2066	2066-2596	2596-3126	3126-3656	3656-4186	4186-4716
Количество коров	23	30	24	14	3	3	2	1

Необходимо найти:

1. среднюю величину удоя по данной выборке;

2. среднее линейное отклонение удоя;

3. среднее квадратическое выборочное отклонение удоя;

4. исправленную выборочную дисперсию;

5. выборочный коэффициент вариации.

Решение. Здесь имеем интервальный сгруппированный ряд. Взяв в качестве представителя каждого интервала его середину, требуемые выборочные характеристики будем находить так же, как и для дискретного ряда. Все необходимые расчеты проведены в таблице 2.6.

1. Средняя величина удоя (по формуле (2.2)) составит:

$$\bar{x} = \frac{1}{100} \cdot \sum_{i=1}^{8} x_i \cdot m_i = \frac{1}{100} \cdot 162080 = 1620{,}8(кг).$$

(Требуемая сумма найдена в пятом столбце табл. 2.6).

2. Среднее линейное отклонение удоя найдем по формуле (2.9):

$$d = \frac{1}{100} \cdot \sum_{i=1}^{8} |x_i - \bar{x}| \cdot m_i = \frac{1}{100} \cdot 61458{,}8 = 614{,}588(кг).$$

(Сумма $\sum_{i=1}^{8} |x_i - \bar{x}| \cdot m_i$ вычислена в столбце 7 табл. 2.6.)

Таблица 2.6

i	Интервалы $x_i < X \le x_{i+1}$	Середина интервала x_i	m_i	$x_i m_i$	$\lvert x_i - \bar{x}\rvert$	$\lvert x_i - \bar{x}\rvert \cdot m_i$	$x_i^2 \cdot m_i$
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	476-1006	741	23	$741 \cdot 23 = 17043$	879,8	20235,5	$741^2 \cdot 23 = ?$
2	1006-1536	1271	30	$1271 \cdot 30 = 38130$	349,8	10494	$1271^2 \cdot 30 = ?$
3	1536-2066	1801	24	$1801 \cdot ? = 43224$	180,2	?	$1801^2 \cdot 24 = ?$
4	2066-2596	2331	14	$? \cdot 14 = 32634$	?	9942,8	$2331^2 \cdot ? = ?$
5	2596-3126	2861	3	?	1240,2	3720,6	$2861^2 \cdot ? = ?$
6	3126-3656	3391	3	?	1770,2	5310,6	?
7	3656-4186	3921	2	?	?	?	?
8	4186-4716	4451	1	?	2830,2	2830,2	?
	$\sum$	–	100	162080	–	61458,8	324620860

3. Среднее квадратическое выборочное отклонение удоя вычислим по формуле (2.18):

$$\sigma = \sqrt{\frac{1}{100}\sum_{i=1}^{8} x_i^2 \cdot m_i - \overline{x}^2} = \sqrt{\frac{1}{100}\cdot 324620860 - (1620{,}8)^2} = \sqrt{619215{,}96} \approx 786{,}9(кг)$$

4. Исправленная выборочная дисперсия (по формуле (2.14)):

$$S^2 = \frac{100}{100-1}\cdot D_в = \frac{100}{99}\cdot \sigma_в{}^2 = \frac{100}{99}\cdot 619215{,}96 \Rightarrow S^2 \approx 625470{,}67.$$

5. Выборочный коэффициент вариации (по формуле (2.19)) равен:

$$V = \frac{\sigma_в}{\overline{x}_в}\cdot 100(\%) = \frac{786{,}9}{1620{,}8}\cdot 100(\%) \approx 48{,}6\%.$$

Вывод: выборка неоднородная.

Ответ. 1. $\overline{x}_в = 1620{,}8$; 2. $d = 614{,}588$ 3. $\sigma_в \approx 786{,}9$;

4. $S^2 \approx 625470{,}67$; 5. $48{,}6\%$.

Задача 2.4. По выборке *В* (*Таблица* 1 *Приложения* 1) определить:

1. смещенную оценку генеральной дисперсии; 2. моду; 3. медиану.

Решение. $\boxed{?}$ Построить соответствующий интервальный ряд.

1. Смещенной оценкой генеральной дисперсии является выборочная дисперсия. Выборочную среднюю и выборочную дисперсию для упрощения счета найдем методом «условного нуля», взяв в качестве $c = 70$ (т.е. середину интервала, соответствующего варианте с наибольшей частотой); $h = 2$ (длина интервала). Соответствующие расчеты проведены в таблице 2.7.

Таблица 2.7.

i	Интервал	Середина интервала x_i	m_i	$\frac{x_i - c}{h}$	$\frac{x_i - c}{h}\cdot m_i$	$\left(\frac{x_i - c}{h}\right)^2$	$\left(\frac{x_i - c}{h}\right)^2 \cdot m_i$	$m_x^{нак}$
1	**2**	**3**	**4**	**5**	**6=5×4**	**7=5^2**	**8=7×4**	**9**
1	59-61	60	1	$\frac{60-70}{2} = -5$	$-5\cdot 1 = -5$	$(-5)^2 = 25$	$25\cdot 1 = 25$	1

2	61-63	62	2	$\frac{62-70}{2}=-4$	$-4\cdot 1=-8$	$(-4)^2=16$	$16\cdot 2=32$	3
3	63-65	64	7	$\frac{64-70}{2}=-3$	$-5\cdot ?=-21$	?	$9\cdot 7=63$	10
4	65-67	66	16	$?=-2$	$?\cdot ?=-32$	?	$?\cdot 16=?$	26
5	67-69	68	27	$?=-1$	?	1	27	53
6	69-71	70	40	$?=0$	?	0	0	93
7	71-73	72	38	$\frac{72-70}{2}=1$	?	1	38	-
8	73-75	74	38	$\frac{74-70}{2}=2$	?	4	152	-
9	75-77	76	18	?	54	9	?	-
10	77-79	78	9	?	36	16	?	-
11	79-81	80	3	?	15	25	75	-
12	81-83	82	1	$\frac{82-70}{2}=6$	6	?	36	-
	$\sum$		200	-	132	-	818	-

По формуле (2.3) получаем:

$$\overline{x}_в = \frac{2}{200}\cdot\sum_{i=1}^{12}\left(\frac{x_i-70}{2}\right)\cdot m_i + 70 = \frac{1}{100}\cdot 132 + 70 = 71{,}32.$$

- Сумма $\sum_{i=1}^{k}\left(\frac{x_i-c}{h}\right)\cdot m_i = \sum_{i=1}^{12}\left(\frac{x_i-70}{2}\right)\cdot m_i$ найдена в столбце 6 табл. 2.7.

Аналогично находим искомую выборочную дисперсию методом «условного нуля» по формуле:

$$D_в = \frac{h^2}{n}\cdot\sum_{i=1}^{k}\left(\frac{x_i-c}{h}\right)^2\cdot m_i - (\overline{x}-c)^2,$$

где x_i, m_i, n, h, c - имеют тот же смысл, что и в формуле (2.3).

Таким образом, получаем:

$$D_в = \frac{2^2}{200}\cdot\sum_{i=1}^{12}\left(\frac{x_i-70}{2}\right)^2\cdot m_i - (71{,}32-70)^2 = 0{,}02\cdot 818 - 1{,}7424 = 14{,}6176.$$

2. Моду (так как ряд интервальный) находим по формуле (2.4):

$$M_0 \approx 69 + 2 \cdot \frac{40-27}{(40-27)+(40-38)} = 69 + 2 \cdot \frac{13}{13+2} = 70{,}7.$$

- Здесь $x_0 = 69$ – начало модального интервала (т. е. интервала с наибольшей частотой). Номер модального интервала $i = 6$, поэтому соответствующие частоты: $m_i = m_6 = 40$; $m_{i-1} = m_5 = 27$; $m_{i+1} = m_7 = 38$.

3. Медиану определяем по формуле (2.6):

$$M_e = 71 + 2 \cdot \frac{\frac{299}{2} - 93}{38} = 71 + 0{,}4 = 71{,}4.$$

- Здесь $x_0 = 71$ – начало медианного интервала; $m_i = m_7 = 38$ – частота медианного интервала; $T_{i-1} = T_6 = 93$ – сумма частот интервалов, предшествующих медианному (найдена в столбце 9 табл. 2.7).

$\boxed{!}$ В столбце накопленных частот суммирование частот произведено только для первых шести интервалов, так как нас интересует здесь только T_6.

Ответ. 1. $D_в = 14{,}6176$; 2. $M_0 = 70{,}7$; 3. $M_e = 71{,}4$.

2.3. Задачи для самостоятельного решения

21. Из генеральной совокупности извлечена выборка объема $n = 50$:

варианта x_i	2	5	7	10
частота m_i	16	12	8	14

Найти: 1. несмещенную оценку:

а) генеральной средней; б) генеральной дисперсии;

2. моду; 3. медиану.

22. Найти выборочную среднюю по данному распределению выборки (методом условных вариант):

x_i	3250	3270	3280
m_i	2	5	3

23. По выборке объема $\mathrm{n} = 81$ найдена смещенная оценка $D_в = 6$ генеральной совокупности. Найти несмещенную оценку дисперсии генеральной совокупности.

24. Для некоторой выборки известно, что выборочная дисперсия $D_в = 5{,}15$, а исправленная $S^2 = 5{,}2015$. Найти объем выборки.

25. Дано распределение выборки объема $\mathrm{n} = 100$:

x_i	340	360	375	380
m_i	20	m_2	18	12

Необходимо найти:

1. величину m_2;

2. моду M_0;

3. выборочную дисперсию $D_в$; 4. исправленную дисперсию S^2;

5. коэффициент вариации

26. Учет удоев на ферме в июне месяце дал следующее распределение:

Удой молока (в литрах)	5-7	7-9	9-11	11-13	13-15	15-17	17-19
Число коров	5	8	15	32	97	10	3

1. Определить средний удой в день, выборочную дисперсию и среднее квадратическое отклонение.

2. Найти моду и медиану (графически и аналитически).

27. Результаты выборочного обследования 100 рабочих крупного завода, проведенного с целью определения времени, затрачиваемого на обработку детали, приведены в таблице:

Время обработки (мин.)	3,6-4,2	4,2-4,8	4,8-5,4	5,4-6,0	6,0-6,6
Число рабочих	14	33	35	12	6

Требуется найти: 1.среднее время обработки детали; 2. среднее линейное отклонение времени обработки детали; 3. среднее квадратическое отклонение; 4. исправленную выборочную дисперсию; 5. выборочный коэффициент вариации.

28. Выборочное обследование 100 промышленных предприятий РФ по среднегодовой численности рабочих приведено в таблице:

Предприятия со среднегодовой численностью рабочих	Число предприятий
до 100	35
100-200	19
200-500	22
500-1000	11
1000-3000	7
3000-10000	3
более 10000	m_7

1. Найти величину m_7. 2. Определить несмещенную оценку:

а) генеральной средней; б) генеральной дисперсии.

---

29. Из генеральной совокупности извлечена выборка объема $\mathrm{n} = 60$:

x_i	1	3	6	26
m_i	8	40	10	2

Найти: 1. несмещенную оценку:

а) генеральной средней;

б) генеральной дисперсии;

2. моду;

3. медиану.

30. Найти выборочную среднюю по данному распределению выборки объема $\mathrm{n} = 20$ (методом условных вариант):

x_i	2560	2600	2620	2650	2700
m_i	2	3	10	4	1

31. По выборке объема $n = 51$ найдена смещенная оценка $D_в = 5$ генеральной дисперсии. Найти несмещенную оценку дисперсии генеральной совокупности.

32. Исправленная дисперсия для выборки объема $n = 21$ составила 6,3. Найти выборочную дисперсию.

33. Дано распределение выборки объема $n = 50$:

x_i	0,1	0,5	0,6	0,8
m_i	5	15	m_3	10

Найти: 1. m_3; 2. моду M_0; 3. выборочную дисперсию $D_в$;

4. исправленную дисперсию S^2; 5. коэффициент вариации.

Указание: Перейти к условным вариантам $u_i = 10x_i$.

34. Валовой сбор пшеницы с трех участков посевной площади составил:

Размер участка (га)	10	20	30
Урожай (ц/га)	15-17	17-19	19-21

Вычислить средней урожай и показатели колеблемости.

35. Группированные данные по запасам обуви в 100 магазинах приведены в следующей таблице:

Интервал, тыс. у. е.	0,2-2,2	2,2-4,2	4,2-6,2	6,2-8,2	8,2-10,2	10,2-12,2
Число магазинов в интервале	70	20	4	3	2	1

Требуется найти: 1) средний запас обуви (в тыс. условных единиц);

2) среднее линейное отклонение запасов обуви;

3) среднее квадратическое отклонение запасов обуви;

4) исправленную выборочную дисперсию;

5) моду и медиану (графически и аналитически).

36. На молочной ферме испытывали эффективность различных кормов. При кормлении коров смесью №1 было получено распределение надоев «А», при кормлении смесью №2 – распределение «В». На основе полученных распределений выяснить, какая смесь кормов лучше для удойности коров.

Надои молока (л)	Количество коров	
	в распределении «А»	в распределении «В»
0-4	3	0
4-8	8	5
8-12	21	9
12-16	11	14
16-20	6	13
20-24	1	8
24-28	0	1
$\sum$	50	50

---

37. Из генеральной совокупности извлечена выборка объема $n = 30$:

x_i	2	4	8	6
m_i	5	13	10	2

Найти: 1. несмещенную оценку:

а) генеральной средней;

б) генеральной дисперсии;

2. моду;

3. медиану.

38. В итоге четырех измерений некоторой величины одним прибором (без систематических ошибок) получены следующие результаты: 8; 9; 11; 12.

Определить:

1. выборочную среднюю результатов измерений;
2. выборочную дисперсию ошибок прибора;
3. исправленную дисперсию ошибок прибора.

39. Найти исправленную выборочную дисперсию по данному распределению выборки:

x_i	23,5	26,1	28,2	30,4
m_i	2	3	4	1

Указание. Перейти к условным вариантам $u_i = 10x_i - 268$.

40. Дано распределение 200 рабочих по месячной заработной плате:

Заработная плата (в руб.)	4000	5000	6000	7000	8000	9000	10000
Количество рабочих	12	21	40	55	38	23	11

Вычислить показатели вариации данного ряда распределения.

41. При обследовании годового надоя коров случайным образом отобрали 307 коров, данные по ним сгруппировали и составили таблицу:

Надои (в кг)	3000-3400	3400-3800	3800-4200	4200-4600	4600-5000
Число коров	43	71	102	64	27

1. Найти выборочное среднее, дисперсию (выборочную и исправленную), средне квадратическое отклонение.
2. Определить моду и медиану (аналитически и графически).

42. Данные об урожайности ржи на различных участках поля сельхозпредприятия приведены в следующей таблице:

Урожайность ржи (ц/га)	9-12	12-15	15-18	18-21	21-24	24-27
Доля участка от общей посевной площади (%)	6	12	33	22	19	8

Необходимо: 1) построить кумулятивный ряд и начертить кумуляту;

2) найти моду и медиану; 3)определить среднюю урожайность;

3) вычислить среднее отклонение урожайности:

а) линейное; б) квадратическое;

4) найти исправленную выборочную дисперсию.

43. Группированная выборка распределения рейтинга успеваемости студентов (в баллах) представлена в таблице:

Успеваемость (баллы)	100-150	150-200	200-250	250-300	300-350	350-400	400-450
Количество студентов	8	15	18	26	16	12	5

Требуется:

1. построить гистограмму и кумуляту выборки;
2. найти смещенную и несмещенную оценки генеральной дисперсии.

44. Имеются следующие данные о жирности молока коров в стаде:

Жирность (в %)	3,5	3,6	3,7	3.8	3,9	4,0	4,1	4,2
Число коров	8	11	28	43	71	52	29	8

1) Преобразовать данный вариационный ряд в интервальный с равными интервалами так, чтобы приведенные в таблице значения признака, были центрами интервалов и построить гистограмму частостей; 2) вычислить коэффициент вариации.

45. Трое рабочих изготавливали одинаковые детали. Каждый из них изготовил по 70 деталей. Для контроля составлены распределения изготовленных каждым рабочим деталей по их размеру. Точность измерения 0,01 мм. Поле допуска (т. е. совокупность допустимых значений размеров деталей) 50,04 – 50,28 мм. Сравнением распределений выяснить, какой рабочий работает лучше.

Размер деталей (мм)	Количество деталей данного размера у		
	I рабочего	II рабочего	III рабочего
50,00-50,04	2	0	3
50,04-50,08	6	2	8
50,08-50,12	9	5	16
50,12-50,16	26	9	9
50,16-50,20	14	14	4
50,20-50,24	7	19	13
50,24-50,28	5	8	11
50,28-50,32	1	5	4
50,32-50,36	0	5	2
50,36-50,40	0	2	0
50,40-50,44	0	1	0
$\sum$	70	70	70

46. Задано распределение первоначальных вариант выборки объема n:

x_i	x_1	x_2	…	x_k
m_i	m_1	m_2	…	m_k

Доказать, что $\overline{x}_в = c + \frac{1}{n} \cdot \sum_{i=1}^{k} m_i u_i$, где условные варианты $u_i = x_i - c$.

47. После обработки 100 деталей определили их размер с точностью до 0,1 мм и составили ряд их распределения по размеру. Затем произвели наладку станка и, обработав такое же количество деталей, составили новый ряд распределения деталей по размеру. Сравнением двух полученных распределений установить, привела ли наладка станка к улучшению его работы, имея в виду, что поле допуска от 19,5 до 20,7 мм.

Размер деталей (мм)	Количество деталей	
	до наладки станка	после наладки станка
19,4-19,6	3	6
19,6-19,8	10	15
19,8-20,0	11	21
20,0-20,2	15	25
20,2-20,4	21	18
20,4-20,6	16	11
20,6-20,8	12	3
20,8-21,0	6	1
21,0-21,2	5	0
21,2-21,4	1	0
$\sum$	100	100

48. По выборке С (*Таблица* 1 *Приложения* 1) вычислить числовые характеристики вариационного ряда:

1.выборочное среднее $\overline{x}_{в}$; 2. выборочную дисперсию $D_{в}$;

3. выборочное среднее квадратическое отклонение $\sigma_{в}$;
4. исправленную выборочную дисперсию s^{2};
5. исправленное среднее квадратическое отклонение s;

4. моду M_0; 5. медиану M_e.

49. 1) Определить, является ли выборка сельхозпредприятий Приморского края, занимающихся производством и реализацией мяса свиней (*таблица* 2 *Приложения* 1) однородной.

2) Найти моду M_0 и медиану M_e данной выборки.

3) Вычислить исправленную выборочную дисперсию s^2 и исправленное среднее квадратическое отклонение s данной выборки.

50. 1) Определить несмещенные оценки генеральной средней и генеральной дисперсии для выборки сельхозпредприятий Приморского края, выращивающих зерновые (*таблица* 3 *Приложения)*

2) Найти моду M_0 и медиану M_e данной выборки.

3) Вычислить выборочный коэффициент вариации.

Тест 2

1. Из следующих числовых характеристик выборки:

1) выборочная средняя; 2) мода; 3) медиана; 4) выборочная дисперсия; 6) размах вариации – показателями колеблемости являются:

а) 1,2,3,4,5,6; б) 4,5,6;

в) 1,2,3; г) 1,4,5.

2. Если x_i $\left(i = \overline{1,n}\right)$ – варианты признака X выборки объема n, то выборочная средняя (простая) находится по формуле:

а) $\overline{x}_в = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i$; б) $\overline{x}_в = \dfrac{1}{n}\sum\limits_{i=1}^{n} x_i^2$;

г) $\overline{x}_в = \dfrac{x_{\max} - x_{\min}}{n}$; г) $\overline{x}_в = \dfrac{1}{n} + \sum\limits_{i=1}^{n} x_i$.

3. Задана группированная выборка распределения случайной величины $У$:

Значения y_i	1,2	1,5	2,0	2,5
Частоты m_i	5	15	20	10

Смещенная дисперсия равна . . .

4. Выборочное среднее $\overline{x}_в = 15$, выборочное среднее квадратическое отклонение $\sigma_в = 3$. Тогда выборочный коэффициент вариации равен:

а) 0,2(%); б) 5(%); в) 500(%); г) 20%

5. Установить соответствие:

Название числовой характеристики выборки.	Расчетная формула.
1) Мода интервального вариационного ряда $M_0 =$	а) $\dfrac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}$;
2) Медиана дискретного вариационного ряда при нечетном объеме выборки n $M_e =$	б) $x_0 + h \cdot \dfrac{\frac{n}{2} - T_{i-1}}{m_i}$, где i - номер медианного интервала, x_0 – начало медианного интервала; h – длина медианного интервала; T_{i-1} – сумма частот; m_i – частота медианного интервала;
3) Медиана дискретного вариационного ряда при четном объеме выборки n $M_e =$	в) $x_{\frac{n+1}{2}}$;
4) Медиана интервального вариационного ряда $M_e =$	г) $x_0 + h \cdot \dfrac{m_i - m_{i-1}}{(m_i - m_{i-1}) + (m_i - m_{i+1})}$, где x_0 – начало модального интервала, i –номер модального интервала; m_i, m_{i-1}, m_{i+1} – частоты.

6. Представителем интервала при вычислении числовых характеристик интервальных вариационных рядов является:

а) середина интервала; б) длина интервала;

в) правый конец интервала; г) левый конец интервала.

7. На рис. 2.4 изображен полигон частот некоторой выборки. Тогда мода M_0 этой выборки равна:

а) $M_0 = 10$; б) $M_0 = 9$; в) $M_0 = 5$ г) $M_0 = 7{,}5$.

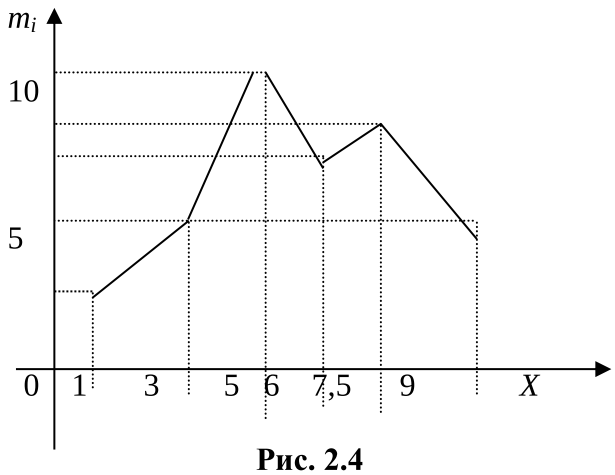


Рис. 2.4

8. Для выборки объема $n = 100$ известно, что $\sum\left|x_i - \bar{x}\right| \cdot m_i = 200$.

Тогда среднее линейное отклонение составит …

9. Для случайно отобранных семи рабочих стаж работы (в годах) оказался равным: 10, 3, 5, 12, 11, 7, 9. Тогда средний стаж равен:

а) 8,14 года; б) 9 лет;

в) 7,15 лет; г) 10 лет.

10. Если n – объем выборки, x_i – варианты изучаемого признака, m_i – его частоты, $\bar{x}_в$ - выборочное среднее, то выборочная дисперсия может быть найдена по формуле:

а) $D_в = \sum x_i^2 \cdot m_i - \bar{x}_в{}^2$;

б) $D_в = \frac{1}{n} \cdot \sum \left(x_i - \bar{x}_в\right)^2 \cdot m_i$;

в) $D_в = \frac{1}{n} \cdot \sum \left(x_i - \bar{x}_в\right)^2 \cdot m_i - \bar{x}_в{}^2$; г) $D_в = \frac{1}{n} \cdot \sum \left(x_i - \bar{x}_в\right)^2$.

11. Оценка некоторого параметра выборки, имеющая наименьшую дисперсию, называется:

а) несмещенной;

б) эффективной;

в) состоятельной;

г) параметрической.

12. Выборочная дисперсия для генеральной дисперсии является:

а) несмещенной оценкой;

б) эффективной оценкой;

в) смещенной оценкой;

г) интервальной оценкой.

13. Из генеральной совокупности извлечена выборка объема $n = 50$:

Значения x_i	2	3	5
Варианты m_i	5	15	30

Тогда несмещенная оценка генеральной средней равна . . .

14. Для некоторой выборки известно, что выборочная дисперсия $D_в = 5{,}15$, а исправленная дисперсия $S^2 = 5{,}17575$. Тогда объем выборки равен:

а) 100;

б) 150;

в) 175;

г) 200.

15. Для сгруппированного интервального вариационного ряда:

Интервалы	10-20	20-30	30-40	40-50	50-60
Варианты m_i	3	5	10	8	6

1) значение выражения равно: $M_0 + M_e = \ldots$ (где M_0 – мода, M_e – медиана);

2) несмещенная оценка генеральной дисперсии равна . . .

Вопросы для самопроверки по главе 2

1. Выборочное среднее (простое и взвешенное).

2. Мода, ее нахождение для дискретного и интервального вариационных рядов.

3. Медиана, её вычисление для дискретного и интервального вариационных рядов.

4. Среднее линейное отклонение (простое и взвешенное).

5. Дисперсия (простая и взвешенная).

6. Среднее квадратическое отклонение (простое и взвешенное).

7. Понятие статистической оценки

8. Точечная оценка: понятие и виды (несмещенная, эффективная, состоятельная).

9. Что является несмещенной оценкой генеральной средней?

10. Что является несмещенной оценкой генеральной дисперсии?

Глава 3. Интервальные оценки числовых характеристик генеральной совокупности

3.1. Необходимый теоретический минимум

I. Основные законы распределения статистических оценок

Распределение статистических оценок в большинстве случаев достаточно точно описывают такие законы распределения, как нормальный закон распределения, χ^2 (хи – квадрат) Пирсона, Стьюдента (t – распределение).

Нормальный закон распределения

Нормальный закон распределения подробно изучен в курсе теории вероятностей. Доказано, что если число наблюдений n велико, то каким бы ни было распределение случайной величины, из которой делается выборка, в силу центральной предельной теоремы выборочное среднее $\overline{X}_в$ подчиняется закону, близкому к нормальному с параметрами: $M\left(\overline{X}_в\right)=a,\; D\left(\overline{X}_в\right)=\frac{\sigma^2}{n},\quad \sigma\left(\overline{X}_в\right)=\frac{\sigma}{\sqrt{n}}$, то есть:

$$X_в \sim N\left(a;\frac{\sigma}{\sqrt{n}}\right). \qquad \textbf{(3.1)}$$

Распределение χ^2 (хи-квадрат)

Определение 3.1. Если $X_1, X_2, \ldots, X_n$ – независимые, нормально распределённые случайные величины с a=M(X_i)=0 и $\sigma=\sigma(X_i)=1$ (i=1,2…n.) т.е. $X_i \sim N(0;1)$, тогда закон распределения суммы квадратов случайных величин: $\chi^2 = X_1^2 + X_2^2 + \ldots + X_n^2$ называется ***законом хи-квадрат*** с n степенями свободы (число степеней свободы равно числу слагаемых).

Из определения 3.1 следует, что случайная величина χ^2 принимает только положительные значения. Математическое ожидание и дисперсия соответственно равны $M\left(\chi_n^{\ 2}\right) = n;\ \ D\left(\chi_n^{\ 2}\right) = 2n.$ Для этого распределения существуют подробные таблицы, которые составлены для $n \le 30$: по заданному уровню значимости α и числу степеней свободы k находят $\chi^2_{\alpha,k}$ из условия $P\left(\chi^2 > \chi^2_{\alpha,k}\right) > \dfrac{\alpha}{2}$. При n>30 распределение случайной величины $X = 2\chi^2 - \sqrt{2n}$ близко к стандартному нормальному закону $N(0;1)$.

Можно показать, что величины $\dfrac{nD_в}{\sigma^2}$, $\dfrac{(n-1)S^2}{\sigma^2}$ имеют распределение χ^2 с числом степеней свободы (n-1). То есть:

$$\frac{nD_в}{\sigma^2} = \frac{(n-1)S^2}{\sigma^2} = \chi_{n-1}^{\ 2}. \qquad \textbf{(3.2)}$$

- Распределением χ^2 можно пользоваться для построения доверительного интервала для неизвестной дисперсии.

Распределение Стьюдента

Определение 3.2. ***Распределением Стьюдента*** (***или t-распределением***) называется распределение случайной величины $t = \dfrac{X_0}{\sqrt{\dfrac{1}{n}\chi^2}}$, где X_0 – случайная величина, распределённая по стандартному нормальному закону $N(0;1)$; χ^2 - независимая от X_0 случайная величина, имеющая распределение хи-квадрат с n степенями свободы.

- При $n \to \infty$ t – распределение приближается к нормальному. Практически при n>30, это распределение можно считать приближённо нормальным. Числовые характеристики: M(t)=0; D(t)=$\frac{n}{n-2}$.

В практических задачах (в частности при построении доверительных интервалов для неизвестных параметров генеральной совокупности) используют случайную величину

$$T = \frac{X_0 - \bar{X}_в}{\sqrt{\sum_{i=1}^{n} \frac{\left(X_i - \bar{X}_в\right)^2}{n}}}.$$

Так как $\bar{X}_в \sim N\left(a; \frac{\sigma}{\sqrt{n}}\right)$; $\sigma_в$ – смещённая, а S –несмещённая оценка для среднеквадратического отклонения можно доказать, что статистики:

$$t = \frac{\sqrt{n}\left(\bar{X}_в - a\right)}{S} = \frac{\sqrt{n-1}\left(\bar{X}_в - a\right)}{\sigma_в} \tag{3.3}$$

имеют распределение Стьюдента с (n –1) степенями свободы.

- Для распределения Стьюдента составлены таблицы (*таблица* 4 *приложения* 2), в которой по числу степенней свободы k и вероятностью γ, находят соответствующее значение t. Для проверки статистических гипотез используют таблицу критических точек распределения Стьюдента (*таблица* 6 *приложения* 2).

II. Понятие доверительного интервала

Если на основании имеющихся данных (выборки из генеральной совокупности) находят оценку $\tilde{\theta}\left(x_1, x_2 .. x_n\right)$ параметра θ генеральной совокупности, то при этом понимают, что величина $\tilde{\theta}$ является лишь приближённым значением неизвестного параметра θ даже в том случае, если эта оценка состоятельная, несмещенная и эффективная. Полученная оценка называется точечной оценкой. Чтобы получить представление о точности и надёжности оценки $\tilde{\theta}$ неизвестного параметра θ, используют *интервальные оценки*.

Определение 3.3. Вероятность γ выполнения неравенства $\left|\tilde{\theta}-\theta\right|<\varepsilon$ называется ***доверительной вероятностью*** или ***надёжностью*** оценки $\tilde{\theta}$, если справедливо следующее:

$$P\left(\left|\tilde{\theta}-\theta\right|<\varepsilon\right)=\gamma \quad \text{или} \quad P\left(\tilde{\theta}-\varepsilon<\theta<\tilde{\theta}+\varepsilon\right)=\gamma.$$

- Оцениваемый параметр $\theta -$ число постоянное, а величина $\tilde{\theta}$ –

случайная, поэтому *доверительная вероятность* γ есть вероятность того, что интервал $\left(\tilde{\theta}-\varepsilon;\, \tilde{\theta}+\varepsilon\right)$ «накрывает» оцениваемый параметр θ.

Определение 3.4. Случайный интервал $\left(\tilde{\theta}-\varepsilon;\, \tilde{\theta}+\varepsilon\right)$, в пределах которого с вероятностью γ находится неизвестный оцениваемый параметр θ, называется ***доверительным интервалом*** I_γ, соответствующим *коэффициенту доверия* γ, т.е.

$$I_\gamma=\left(\tilde{\theta}-\varepsilon;\, \tilde{\theta}+\varepsilon\right).$$

- Надёжность оценки γ можно задавать заранее (обычно $\gamma=0{,}95$; $0{,}99$; $0{,}999$), тогда зная закон распределения изучаемой величины, находят доверительный интервал. Решают и обратную задачу, когда по заданному доверительному интервалу $I\gamma$ находят соответствующую надёжность оценки.

Определение 3.5. Число $\alpha = (1-\gamma)$ называется ***уровнем значимости*** или ***вероятностью ошибки***.

- Уровень значимости задаётся заранее в зависимости от конкретного случая.

III. Построение доверительного интервала для генеральной средней

Доверительный интервал для генеральной средней при известном среднеквадратическом отклонении генеральной совокупности σ

Предположим, что задана доверительная вероятность γ и объём выборки велик $(n \to \infty)$, тогда выборочная средняя $\overline{X}_в$ приближённо имеет нормальный закон распределения $\overline{X}_в \sim N\left(a; \dfrac{\sigma}{\sqrt{n}}\right)$ (см.(3.1)), где $a = M(X) = X_{ген.}$ (σ известно).

Из свойств нормального распределения имеем:

$$P\left(\left|\overline{X}_в - X_{ген.}\right| < \varepsilon\right) = 2\Phi\left(\frac{\varepsilon\sqrt{n}}{\sigma}\right),$$

если обозначим $t = \dfrac{\varepsilon\sqrt{n}}{\sigma}$, то $P(\left|\overline{X}_в - X_{ген.}\right| < \varepsilon) = 2\Phi(t) = \gamma$,

где ε – ***предельная ошибка выборки***, которая находится по формуле:

$$\varepsilon = \frac{t\sigma}{\sqrt{n}}. \qquad \textbf{(3.4)}$$

Соответственно ***доверительный интервал*** для генеральной средней при ***известном*** значении σ задается интервалом:

$$X_{ген.} \in \left(\overline{X}_в - \frac{t\sigma}{\sqrt{n}};\ \overline{X}_в + \frac{t\sigma}{\sqrt{n}} \right). \qquad \textbf{(3.5)}$$

Значение t по заданной надежности γ находят с помощью таблицы значений *интегральной функции Лапласа* (*таблица* 3 *приложения* 2) из следующего соотношения:

$$2\Phi(t) = \gamma. \qquad \textbf{(3.6)}$$

Например, для надежности $\gamma = 0{,}95 \Rightarrow t = 1{,}96.$

*Пример 1***.** Некоторый количественный признак X распределён нормально в генеральной совокупности, известно среднее квадратическое отклонение $\sigma = 5$. По данной выборке объёма n=100, $\overline{X}_в = 12$ найти доверительный интервал для генеральной средней с надёжностью 0,99.

Решение. По условию задачи: n=100, $\sigma = 5,\ \gamma = 0{,}99,\ \ \overline{X}_в = 12$.

Найдём значение t, соответствующее доверительной вероятности γ, из условия (3.6) и затем используя *таблицу* 3 *Приложения* 2:

$$2\Phi(t) = 0{,}99 \Rightarrow \Phi(t) = 0{,}495 \Rightarrow t = 2{,}58\,.$$

Вычисляем предельную ошибку выборки ε по формуле (3.4):

$$\varepsilon = \frac{2{,}58 \cdot 5}{10} = 1{,}29 \approx 1{,}3.$$

Теперь записываем доверительный интервал для генеральной средней (см.(3.5)):

$$12 - 1{,}3 < X_{ген.} < 12 + 1{,}3$$

$$\text{или} \quad 10{,}7 < X_{ген.} < 13{,}3.$$

Доверительный интервал для генеральной средней при неизвестном среднеквадратическом отклонении генеральной совокупности σ

Пусть количественный признак X в генеральной совокупности распределён нормально, объём выборки велик (n>30). В этом случае значение σ в формулах (3.4) и (3.5) заменяют исправленным выборочным средним квадратическим отклонением S, являющимся несмещённой оценкой $\sigma(X)$.

Итак, ***доверительный интервал*** для генеральной средней при ***неизвестном*** σ (n >30) задается интервалом:

$$X_{ген.} \in \left(\overline{X}_в - \frac{tS}{\sqrt{n}};\ \overline{X}_в + \frac{tS}{\sqrt{n}} \right). \qquad \textbf{(3.7)}$$

Соответствующая ***предельная ошибка*** равна $\varepsilon = \dfrac{t \cdot S}{\sqrt{n}}$, где исправленное среднее квадратическое отклонение $S = \sqrt{\dfrac{n}{n-1}} \cdot \sigma_в$, а значение t находят из соотношения (3.6).

- На практике при больших n выборочное ($\sigma_в$) и исправленное (S) средние квадратические отклонения отличаются мало, поэтому исправленным отклонением пользуются при n<30.

*Пример 2***.** Взвешивание 50 случайно отобранных поросят при рождении дало следующие результаты: $\overline{X}_в = 1200г,\ \sigma_в = 108г$. Найти интервальную оценку генеральной средней при доверительной вероятности 0,95.

Решение. Используем формулу (3.7), для этого необходимо найти S и t. Несмещённая оценка для среднего квадратического отклонения

$$S = \sqrt{\frac{n}{n-1}} \cdot \sigma_в \Rightarrow S = \sqrt{\frac{50}{49}} \cdot 108 = 109{,}096 \approx 109{,}1.$$

Значение t находим из соотношения (3.6), по *таблице* 3 (*Приложение 2*):

$$2\Phi(t) = 0{,}95 \;\Rightarrow\; \Phi(t) = 0{,}475 \;\Rightarrow\; t = 1{,}96.$$

Предельная ошибка выборки $\varepsilon = \dfrac{t \cdot S}{\sqrt{n}} \;\Rightarrow \varepsilon = \dfrac{1{,}96 \cdot 109{,}1}{\sqrt{50}} = 30{,}24 \approx 30$.

Интервальная оценка генеральной средней следующая:

$$1200 - 30 < X_{ген.} < 1200 + 30 \quad \text{или} \quad 1170 < X_{ген.} < 1230.$$

Получили, что вес новорожденных поросят изменяется в пределах от 1170 *г* до 1230 *г*.

Оценка генеральной средней по малой выборке

Предположим, что в генеральной совокупности рассматриваемый признак имеет нормальное распределение. В случае большой выборки (n>30), доверительные интервалы построены (см. (3.5) и (3.7)). Можно заметить, что чем меньше n, тем шире интервал т.е. величина интервала зависит от объёма выборки. Но сделать большую выборку иногда не удаётся и это не всегда целесообразно. Если известна генеральная дисперсия σ^2, то доверительный интервал можно построить аналогично изложенному и при малых значениях n, учитывая, что статистика $T = \dfrac{\overline{X}_в - X_{ген.}}{\sigma} \cdot \sqrt{n}$

имеет нормальное распределение $N(0;1)$.

На практике генеральная дисперсия и оцениваемая генеральная средняя неизвестны, поэтому σ^2 заменяют её наилучшей оценкой, которой является исправленная выборочная дисперсия S^2.

Тогда статистика: $\tau = \dfrac{\overline{X}_в - X_{ген.}}{S} \cdot \sqrt{n}$

имеет *t-распределение Стьюдента* с $k = n$–1 степенями свободы (см. (3.3)).

Если $P\left(\left|\overline{X}_в - X_{ген.}\right| \le \varepsilon_{м.в.}\right) = \gamma$, где γ -доверительная вероятность, то предельная ошибка малой выборки составит:

$$\varepsilon_{м.в.} = \frac{t_{\gamma,n-1} \cdot S}{\sqrt{n}}. \qquad \textbf{(3.8)}$$

Величину $t_{\gamma,n-1}$ при заданной доверительной вероятности γ и числе степеней свободы $k = n - 1$ находят по таблице *значений критерия Стьюдента* (*табл.* 4 *приложения 2*).

Тогда доверительный интервал для генеральной средней (малая выборка n<30) имеет вид:

$$\overline{X}_в - \frac{t_{\gamma,n-1} \cdot S}{\sqrt{n}} \le X_{ген.} \le \overline{X}_{в.} + \frac{t_{\gamma,n-1} \cdot S}{\sqrt{n}}. \qquad \textbf{(3.9)}$$

Пример 3. По данным выборки объёма n=16 из генеральной совокупности нормально распределённого количественного признака X найдены средняя выборочная $\overline{X}_в = 42{,}8$ и исправленное среднее квадратическое отклонение S=8. Найти доверительный интервал для генеральной средней с надёжностью 0,99. *Решение*. Для вычисления предельной ошибки ε (см.(3.8)) найдём значение $t_{\gamma,n-1}$. Из условия примера следует, что $\gamma = 0{,}99,\ \ k = n - 1 = 16 - 1 = 15.$

По *таблице значений критерия Стьюдента таб.* 4 (*Приложение 2*) определяем значение t = 2,95. Тогда ошибка выборки равна:

$$\varepsilon = \frac{2{,}95 \cdot 8}{\sqrt{16}} = 5{,}9.$$

Соответственно, доверительный интервал имеет вид:

$$42{,}8 - 5{,}9 < X_{ген.} < 42{,}8 + 5{,}9 \quad \text{или} \quad 36{,}9 < X_{ген.} < 48{,}7.$$

Оценка требуемого объёма выборки

Формулы доверительного интервала позволяют также решить следующую задачу: каким должен быть объём выборки *n*, чтобы с надёжностью γ точность оценки, полученной по ней для $X_{ген.}$, не превосходила заданного значения ε, то есть выполнялось неравенство: $\left|\overline{X}_в - X_{ген.}\right| < \varepsilon$.

Из формулы (3.4) выразим *n*, получив следующее выражение:

$$n = \left(\frac{t \cdot \sigma}{\varepsilon}\right)^2. \qquad (3.10)$$

Пример 4. Найти минимальный объём выборки, для которой с надёжностью 0,925 точность оценки генеральной средней нормально распределённой генеральной совокупности по выборочной средней равна 0,2, если известно, что среднеквадратическое отклонение генеральной совокупности $\sigma = 1{,}5$.

Решение. Дано: $\gamma = 0{,}925$; $\varepsilon = 0{,}2$; $\sigma = 1{,}5$. По значению γ найдём *t*, из соотношения: $2\Phi(t) = \gamma$ (см.(3.6)). В нашем примере $2\Phi(t) = 0{,}925$ $\Rightarrow \Phi(t) = 0{,}4625$. По *таблице* 3 (*Приложение* 2) находим значение t = 1,78. Вычисляем теперь n по формуле (10.10): $n = \left(\frac{1{,}78 \cdot 1{,}5}{0{,}2}\right)^2 = 178{,}22$.

Так как минимальный объём выборки есть целое число, то, «округляя», получаем: n=179.

IV. Построение доверительного интервала для генеральной дисперсии

Предполагаем, что распределение признака в генеральной совокупности – нормальное. Тогда случайная величина (см. (3.2))

$$\frac{n\cdot D_в}{\sigma^2}=\frac{(n-1)S^2\cdot}{\sigma^2}=\chi_{n-1}{}^2$$

имеет распределение χ^2 с $k = n–1$ степенями свободы. Если задан уровень доверия γ, то для случайной величины $\chi_{n-1}{}^2$ по таблицам распределения χ^2 надо найти интервал, в который величина, имеющая такое же распределение, попадает с вероятностью γ то есть:

$$P\left(\chi_1{}^2<\chi_{n-1}{}^2<\chi_2{}^2\right)=\gamma \quad \text{или} \quad P\left(\chi_1{}^2<\frac{n\cdot D_в}{\sigma^2}<\chi_2{}^2\right)=\gamma \Rightarrow$$

$$P\left(\frac{n\cdot D_в}{\chi_2{}^2}<\sigma^2<\frac{n\cdot D_в}{\chi_1{}^2}\right)=\gamma\ .$$

Следовательно, ***доверительные интервалы для генеральной дисперсии*** можно задать в одном из следующих двух видов:

$$\frac{n\cdot D_в}{\chi_2{}^2}<\sigma^2<\frac{n\cdot D_в}{\chi_1{}^2}, \qquad (3.11)$$

$$\frac{(n-1)\cdot S^2}{\chi_2{}^2}<\sigma^2<\frac{(n-1)\cdot S^2}{\chi_1{}^2}. \qquad (3.12)$$

- $D_в$ – выборочная дисперсия, S^2 – несмещённая, исправленная оценка для дисперсии.

Соответственно для ***среднеквадратического отклонения*** получим ***доверительные интервалы***, извлекая корни из границ интервалов для генеральной дисперсии:

$$\sigma_в \cdot \sqrt{\frac{n}{\chi_2{}^2}} < \sigma < \sigma_в \cdot \sqrt{\frac{n}{\chi_1{}^2}}, \tag{3.13}$$

$$S \cdot \sqrt{\frac{n-1}{\chi_2{}^2}} < \sigma < S \cdot \sqrt{\frac{n-1}{\chi_1{}^2}}. \tag{3.14}$$

Обычно принято пользоваться несмещённой оценкой для дисперсии S^2 и для среднеквадратического отклонения *S*, т.е. формулами (3.12) и (3.14). Находят $\chi_2{}^2$ и $\chi_1{}^2$ по *таблицам значений* $\chi_{\alpha,k}{}^2$ *критерия Пирсона* (*табл*. 5 *приложения* 2), где число степеней свободы $k = n - 1$ (*n* – объём выборки).

Величины $\chi_2{}^2$ и $\chi_1{}^2$ – критические точки для уровней значимости $\frac{\alpha}{2}$ и $\left(1-\frac{\alpha}{2}\right)$, где $\alpha = 1-\gamma$. Значит, можно записать следующее:

$$\chi_2{}^2 = \chi_{\frac{\alpha}{2},n-1}{}^2; \qquad \chi_1{}^2 = \chi_{\left(1-\frac{\alpha}{2}\right),n-1}{}^2.$$

Значения $\chi_1{}^2$ и $\chi_2{}^2$ определяются неоднозначно при одном и том же значении γ, так как распределение χ^2 не является симметричным. При использовании таблиц χ^2 необходимо учесть:

$$P\left(\chi^2 > \chi_1{}^2\right) = 1-\frac{\alpha}{2}; \qquad P\left(\chi^2 > \chi_2{}^2\right) = \frac{\alpha}{2}.$$

- Таблицы $P\left(\chi^2 > \chi_{\alpha,k}{}^2\right)$ составлены при числе степеней свободы $1 < k < 30$. При $k > 30$ можно считать, что случайная величина $\left(\sqrt{2\chi^2} - \sqrt{2k-1}\right)$ имеет стандартное нормальное распределение $N(0;1)$. Поэтому для определения $\chi_1{}^2$ и $\chi_2{}^2$ следует записать, что $P\left(\left|\sqrt{2\chi^2} - \sqrt{2k-1}\right| < t\right) = 2\Phi(t) = \gamma.$ При расчете доверительного интервала для генеральной дисперсии и генерального среднего квадратического отклонения будем полагать, что:

$$\chi_1{}^2 = \frac{1}{2}\left(\sqrt{2k-1} - t\right)^2; \quad \chi_2{}^2 = \frac{1}{2}\left(\sqrt{2k-1} + t\right)^2,$$

где величину t находим по табл. 3 Приложения 2 из соотношения

$$\Phi(t) = \frac{\gamma}{2}.$$

Найденные значения подставляем в нужные формулы (3.11)–(3.14).

- Ещё один способ выбора интервала позволяет упростить его нахождение. Выбирают интервал, симметричный относительно значения S. Интервал имеет вид:

$$\left[\begin{array}{ll} S\cdot(1-q) < \sigma < S\cdot(1+q) & при\ q < 1; \\ 0 < \sigma < S\cdot(1+q) & при\ q > 1. \end{array}\right. \quad \textbf{(3.15)}$$

Для значений q составлены таблицы (*табл.* 7 *приложения* 2), с помощью которых по известному n (объёму выборки) и доверительной вероятности γ определяют q. Но не в каждом учебнике есть такая таблица, а таблица распределения χ^2 имеется в любом учебнике по теории вероятностей и математической статистике, поэтому здесь и рассмотрены два способа нахождения доверительного интервала для генеральной дисперсии и среднеквадратического отклонения.

Пример 5. Пусть признак X в генеральной совокупности имеет нормальное распределение. Производится выборка, объём которой n=26. Найдено исправленное среднее квадратическое отклонение S=1,8. Найти доверительный интервал, покрывающий среднее квадратическое отклонение σ с надёжностью $\gamma = 0{,}9$.

Решение. При решении задачи используем формулу (3.14). Найдём уровни значимости для определения χ_1^2 и χ_2^2, $\gamma = 0{,}9 \Rightarrow \alpha = 1 - \gamma = 0{,}1$; $\frac{\alpha}{2} = 0{,}05$, $1 - \frac{\alpha}{2} = 0{,}95$, число степеней свободы $k = n - 1 = 25$. Тогда по *таблице* 5 *Приложения* 2 находим:

$$\chi_2^2 = \chi^2_{0,05,25} = 37{,}7; \quad \chi_1^2 = \chi^2_{0,95,25} = 14{,}6 \quad .$$

Искомый доверительный интервал имеет вид:

$$1{,}8 \cdot \sqrt{\frac{25}{37{,}7}} < \sigma < 1{,}8 \cdot \sqrt{\frac{25}{14{,}6}} \quad \text{или} \quad 1{,}47 < \sigma < 2{,}36 .$$

V. Интервальная оценка вероятности (генеральной доли) по относительной частоте

Пусть производятся независимые испытания с неизвестной вероятностью p появления события A в каждом испытании. Требуется оценить неизвестную вероятность p по относительной частоте, то есть найти её точечную и интервальную оценки.

В качестве точечной оценки неизвестной вероятности принимают относительную частоту $\tilde{p} = \frac{m}{n}$. Эта оценка несмещённая, так как

$M\left(\frac{m}{n}\right) = p$. Найдём дисперсию оценки с учётом того, что $D(m)=npq$, где $q = 1 - p$:

$$D(\tilde{p}) = D\left(\frac{m}{n}\right) = \frac{1}{n^2} D(m) = \frac{npq}{n^2} = \frac{pq}{n},$$

тогда среднее квадратическое оценки соответственно равно:

$$\sigma(\tilde{p}) = \sigma\left(\frac{m}{n}\right) = \sqrt{\frac{pq}{n}}.$$

Если n достаточно велико и вероятность p не очень близка к нулю и единице, то можно считать, что относительная частота имеет приближённо нормальное распределение $N(a;\sigma)$, причём $a = M(\tilde{p}) = p$. Строим доверительный интервал (p_1; p_2) который с надёжностью γ покрывает оцениваемый параметр p:

$$P(|\tilde{p} - p| < \varepsilon) = 2\Phi\left(\frac{\varepsilon}{\sigma(\tilde{p})}\right) = 2\Phi\left(\frac{\varepsilon \cdot \sqrt{n}}{\sqrt{pq}}\right) = 2\Phi(t) = \gamma,$$

где $t = \dfrac{\varepsilon\sqrt{n}}{\sqrt{pq}} \Rightarrow \varepsilon = t \cdot \sqrt{\dfrac{pq}{n}}$ *– предельная ошибка.*

Тогда имеем: $P\left(|\tilde{p} - p| < t \cdot \sqrt{\frac{pq}{n}}\right) = \gamma$. Заменив $q=1-p$, получим, что с вероятностью $\gamma = 2\Phi(t)$ выполняется неравенство:

$$\tilde{p} - t \cdot \sqrt{\frac{p(1-p)}{n}} < p < \tilde{p} + t \cdot \sqrt{\frac{p(1-p)}{n}}. \qquad \textbf{(3.16)}$$

Границы интервала зависят от неизвестной вероятности p, поэтому можно решить это неравенство относительно p и найти границы доверительного интервала $p \in (p_1; p_2)$:

$$p_1 = \frac{n}{t^2+n}\left[\tilde{p} + \frac{t^2}{2n} - t \cdot \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n} + \left(\frac{t}{2n}\right)^2}\right];$$

$$p_2 = \frac{n}{t^2+n}\left[\tilde{p} + \frac{t^2}{2n} + t \cdot \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n} + \left(\frac{t}{2n}\right)^2}\right]. \qquad (3.17)$$

Можно воспользоваться тем, что при больших n неизвестную вероятность p заменяют её эмпирическим значением $\tilde{p} = \frac{m}{n}$ и тогда интервал будет иметь вид:

$$\tilde{p} - t \cdot \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}} < p < \tilde{p} + t \cdot \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}. \qquad \textbf{(3.18)}$$

Формула доверительного интервала для p позволяет решить также следующую задачу: каким должен быть объём выборки n, чтобы с надёжностью γ точность оценки, полученной по ней для p не превосходила заданного значения ε, т.е. выполнялось неравенство

$$|\tilde{p} - p| < \varepsilon.$$

Значение n находят из уравнения:

$$t \cdot \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}} = \varepsilon \Rightarrow \qquad n = \frac{t^2}{\varepsilon^2} \cdot \tilde{p}(1-\tilde{p}). \qquad \textbf{(3.19)}$$

- В формуле (3.19) величина $\tilde{p}(1-\tilde{p})=\tilde{p}\tilde{q}$ имеет наибольшее значение, равное 0,25, поэтому если $\tilde{p}$ неизвестно, то для *n* имеем оценку: $n \le \dfrac{t^2}{4\varepsilon^2}$.

- Для оценки генеральной доли *p* нормально распределённого количественного признака по выборочной доле $\tilde{p}=\dfrac{m}{n}$ при малом объёме выборки (*n*<30), формула примет вид:

$$P\left(\tilde{p}-t\cdot\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}<p<\tilde{p}+t\cdot\sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}}\right)=2S(t)=\gamma,$$

где *t* определяют по доверительной вероятности γ и числу степеней свободы $k = n - 1$ с помощью *таблиц значений критерия Стьюдента* (*табл*. 4 *приложения* 2).

Пример 6. Произведено 250 независимых испытаний, в каждом из испытаний вероятность появления события *A* равна *p* .Событие *A* появилось в 32 испытаниях. Найти доверительный интервал, покрывающий неизвестную вероятность с надёжностью 0,99.

Решение. По условию $n=250,\ \ m=32,\ \ \gamma=0{,}99.$ Используем формулу (3.18). Относительная частота появления события *A*: $\tilde{p}=\dfrac{m}{n}=\dfrac{32}{250}=0{,}128$. Для вычисления границ доверительного интервала найдём значение *t* из условия $2\Phi(t)=\gamma.$ В нашей задаче: $2\Phi(t)=0{,}99 \ \Rightarrow\ \Phi(t)=0{,}495 \ \Rightarrow t=2{,}58$ (см. *таблицу* 3 *приложения* 2). Находим границы:

$$p_1 = 0{,}128 - 2{,}58 \cdot \sqrt{\frac{0{,}128 \cdot (1 - 0{,}128)}{250}} = 0{,}0735;$$

$$p_2 = 0{,}128 + 2{,}58 \cdot \sqrt{\frac{0{,}128 \cdot (1 - 0{,}128)}{250}} = 0{,}1825.$$

Искомый доверительный интервал для неизвестной вероятности (генеральной доли) p имеет вид: $0{,}0735 < p < 0{,}1825$.

3.2. Примеры решения типовых задач

Задача 3.1. Исследовалось время безотказной работы 56 лазерных принтеров. Из априорных наблюдений известно, что среднеквадратическое отклонение времени безотказной работы $\sigma = 50$ часов. По результатам исследований получено среднее время безотказной работы $\overline{X}_в = 1500$ часов. Построить: 1. 90%- ый доверительный интервал для среднего времени безотказной работы; 2. 99% - ый доверительный интервал для среднего времени безотказной работы.

Решение.

Для построения интервала необходимо найти предельную ошибку выборки ε по формуле (3.4) при n=56, σ=50, $\gamma_1 = 90\% = 0{,}9$ *и* $\gamma_2 = 99\% = 0{,}99$ - доверительные вероятности. Из соотношения (3.6) и, используя *таблицу* 3 (*Приложение* 2), находим значения t, соответствующие данным вероятностям:

1. $2\Phi(t) = 0{,}9 \Rightarrow \Phi(t) = 0{,}45 \Rightarrow t = 1{,}64.$ Тогда предельная ошибка:

$\varepsilon = \dfrac{1{,}64 \cdot 50}{\sqrt{56}} = 10{,}95 \approx 11$. Искомый доверительный интервал:

$$1500 - 11 \le X_{ген.} \le 1500 + 11 \text{ или } 1489 \le X_{ген.} \le 1511.$$

2. Найдем значение t, используя условие:

$$2\Phi(t) = 0{,}99 \Rightarrow \Phi(t) = 0{,}495 \Rightarrow t = 2{,}58.$$

Тогда предельная ошибка: $\varepsilon = \dfrac{2{,}58 \cdot 50}{\sqrt{56}} = 17{,}23 \approx 17$.

Искомый доверительный интервал:

$1500 - 17 \le X_{ген.} \le 1500 + 17$ или $1483 \le X_{ген.} \le 1517$.

- Как видим, во втором случае интервал получился шире, т.е. величина уровня доверия влияет на ширину интервала: чем больше уровень доверия, тем шире интервал.

Ответ. 1) $1489 \le X_{ген.} \le 1511$; 2) $1483 \le X_{ген.} \le 1517$.

Задача 3.2. Решить *задачу 3.1*, при условии, что априорных сведений о среднеквадратическом отклонении времени безотказной работы нет. Известна лишь оценка $\sigma_в = 59\ (ч.)$, полученная по результатам наблюдений.

Решение.

По условию: n = 56, $\overline{X}_в = 1500,\ \sigma_в = 59.$

1. Пусть $\gamma = 90\%$, t=1,64. Найдем предельную ошибку выборки ε, для этого предварительно вычислим значение S:

$$S = \sqrt{\frac{n}{n-1}} \cdot \sigma_в = \sqrt{\frac{56}{55}} \cdot 59 = 59{,}53;\quad \varepsilon = \frac{t \cdot S}{\sqrt{n}} = \frac{1{,}64 \cdot 59{,}53}{\sqrt{56}} = 13{,}05 \approx 13.$$

Тогда доверительный интервал имеет вид (см. (3.7)):

$1500 - 13 \le X_{ген.} \le 1500 + 13$ или $1487 \le X_{ген.} \le 1513$.

2. При $\gamma =$ 99% , t = 2,58. Тогда предельная ошибка выборки равна:

$$\varepsilon = \frac{2{,}58 \cdot 59{,}53}{\sqrt{56}} = 20{,}53 \approx 21;$$

Доверительный интервал имеет вид (см. выражение $\boxed{\textbf{?}}$):

$$1479 \le X_{ген.} \le 1521.$$

? Проверьте как будут отличаться результаты, если при вычислении ε вместо S подставить $\sigma_в$. Сделайте вывод.

Ответ. 1. $1487 \le X_{ген.} \le 1513$;

2. $1479 \le X_{ген.} \le 1521.$

Задача 3.3. На ферме испытывалось влияние витаминов на прибавку в весе телят. Для этой цели было осмотрено 20 телят одного возраста. Средний вес их оказался равным 340 кг, а исправленное среднеквадратическое отклонение 15 кг. Найти доверительный интервал для генеральной средней с надёжностью 0,95.

Решение. В соответствии с условием задачи $\overline{X}_в = 340,\ S = 15,\ \gamma = 0{,}95,\ \ n = 20.$ Объём выборки n<30, поэтому при расчёте используем распределение Стьюдента. Находим значение $t_{\gamma,n-1}$. В нашем случае $\gamma = 0{,}95$ и число степеней свободы k = 20 –1=19. Тогда по *табл*. 4 *приложения* 2 имеем $t_{\gamma,n-1} = 2{,}09$. Вычислим теперь предельную ошибку выборки (по формуле (3.8)): $\varepsilon = \frac{2{,}09 \cdot 15}{\sqrt{20}} = 7{,}01 \approx 7$. Получим доверительный интервал:

$340 - 7 \le X_{ген.} \le 340 + 7$ или $333 \le X_{ген.} \le 347$.

Ответ. $333 \le X_{ген.} \le 347$.

Задача 3.4. При изучении длины листьев садовой земляники сделана выборка. Среднее квадратическое исправленное отклонение отдельного измерения S=1,32см. С вероятностью 0,95 определите такое минимальное число измерений, чтобы отклонение выборочной средней от математического ожидания не превышало 0,06.

Решение. По условию задачи: S=1,32, $\gamma = 0{,}95, \quad \varepsilon = 0{,}06.$

Для нахождения n- объёма выборки используем формулу (3.10), заменив σ на S, получим $n = \left(\dfrac{t \cdot S}{\varepsilon}\right)^2$. Значение t находим из условия $2\Phi(t) = 0{,}95 \Rightarrow \Phi(t) = 0{,}475 \Rightarrow t = 1{,}96$ (табл. , прил.2) $\boxed{?}$

$n = \left(\dfrac{1{,}96 \cdot 1{,}32}{0{,}06}\right)^2 = 1859{,}33$. Значит, минимальное число измерений n=1860.

Ответ. 1860.

Задача 3.5. Из вкладчиков банка было отобрано 400. Средний размер вклада в выборке составил 8000 рублей, а среднее квадратическое отклонение 2500 рублей. Какова вероятность того, что средний размер вклада случайно выбранного вкладчика отличается от его среднего размера в выборке не более, чем на 100 рублей (по абсолютной величине), запишите доверительный интервал, соответствующий этой вероятности.

Решение.

По условию: $n = 400, \ \overline{X}_в = 8000, \ \sigma_в = 2500, \ \varepsilon = 100.$

Так как объём выборки велик, то можно считать, что $\sigma_в = S$. Из формулы для нахождения ε найдём значение t:

$$\varepsilon = \frac{t \cdot S}{\sqrt{n}} \ \Rightarrow \ t = \frac{\varepsilon \cdot \sqrt{n}}{S} \ ;$$

В нашей задаче $t = \dfrac{100 \cdot \sqrt{400}}{2500} = 0{,}8$, подставим данное значение в выражение $2\Phi(t) = \gamma$, получим $2\Phi(0{,}8) = \gamma \ \Rightarrow \ \gamma = 0{,}5762 \approx 0{,}58$.

(Значение $\Phi(0{,}8) = 0{,}2881$ найдено по табл.3 приложения 2.)

Можно записать ответ в виде: $P\left(\left|\overline{X}_{в} - X_{ген.}\right| \le 100\right) = 0{,}58$.

Соответствующий доверительный интервал: $7900 \le X_{ген.} \le 8100$.

Ответ. $7900 \le X_{ген.} \le 8100$.

Задача 3.6. В *задаче 3.3* определить доверительный интервал для среднего квадратического отклонения с надёжностью 0,95 (двумя способами).

Решение.

По условию: n=20, S=15, $\gamma = 0{,}95$.

1. Выберем более простой способ решения (по формуле (3.15)). По *таблице* 7(*приложение 2*) для n=20 и γ =0,95 найдём значение q=0,37.Подставим найденное значение в формулу (3.15):

$$15(1-0{,}37) < \sigma < 15(1+0{,}37) \Rightarrow 9{,}45 < \sigma < 20{,}55 .$$

Округляя до десятых получим: $\underline{9{,}5 < \sigma < 20{,}6.}$

2. Решим задачу, используя таблицы распределения χ^2, а доверительный интервал найдем по формуле (3.14).

Так как $\gamma = 0{,}95$, то $\alpha = 1-\gamma = 0{,}05$. Число степеней свободы k=n–1=19, уровни значимости для $\chi_2^{\,2}$ и $\chi_1^{\,2}$ соответственно следующие: $\frac{\alpha}{2} = 0{,}025$ и $1-\frac{\alpha}{2} = 0{,}975$. По табл. 5 (приложение 2):

$$\chi_2^{\,2} = \chi^2_{0{,}025,19} = 32{,}9; \quad \chi_1^{\,2} = \chi^2_{0{,}975,19} = 8{,}91.$$

Подставляем найденные значения в формулу (3.14):

$$15\cdot\sqrt{\frac{19}{32{,}9}} < \sigma < 15\cdot\sqrt{\frac{19}{8{,}91}} \Rightarrow \underline{11{,}4 < \sigma < 21{,}9}.$$

- Приведённый пример (судя по ответам) показывает зависимость интервала от способа построения,

Ответ. 1. $9{,}5 < \sigma < 20{,}6.$

2. $11{,}4 < \sigma < 21{,}9$

Задача 3.7. Для оценки числа безработных среди рабочих одного из районов города отобрали 400 человек рабочих специальностей, 25 из них оказались безработными. Используя 95% доверительный интервал, оценить истинные размеры безработицы среди рабочих этого района.

Решение.

По условию: n=400, m=25, $\gamma = 0{,}95$. Так как объём выборки велик, то для решения используем формулу (3.18). Найдём относительную частоту $\tilde{p} = \frac{m}{n} = \frac{25}{400} = 0{,}0625$. По доверительной вероятности из условия $2\Phi(t) = \gamma = 0{,}95$ (по табл. 3 приложения 2) находим t=1,96. Вычисляем предельную ошибку:

$$\varepsilon = t \cdot \sqrt{\frac{\tilde{p}(1-\tilde{p})}{n}} = 1{,}96 \cdot \sqrt{\frac{0{,}0625(1-0{,}0625)}{400}} = 0{,}0237.$$

Доверительный интервал будет иметь вид:

0,0625-0,0237<p<0,0625+0,0237 или <u>0,0388<p<0,0862 .</u>

Ответ. Истинный размер безработицы находится приближённо в пределах от 4% до 9%.

Задача 3.8. Выборочное обследование показало, что доля покупателей, предпочитающих новую модификацию товара *A*, составляет 60% от общего числа покупателей данного товара.

1. Каким должен быть объём выборки, чтобы можно было получить оценку генеральной доли с точностью 0,05 при доверительной вероятности 0,9?

2. Решить эту же задачу при условии, что процент покупателей, предпочитающих данный товар, неизвестен.

Решение.

1.По условию: $\tilde{p} = 60\% = 0{,}6,\ \varepsilon = 0{,}05,\quad \gamma = 0{,}9.$ Для определения объёма выборки *n* используем формулу (3.19) . По данной доверительной вероятности найдём значение *t* из соотношения $2\Phi(t) = 0{,}9 \Rightarrow$ $\Phi(t) = 0{,}45$. По *таблице* 3 (*приложение2*) получим: *t* =1,64. Подставляем данные и найденные значения в формулу:

$n = \dfrac{t^2}{\varepsilon^2} \cdot \tilde{p}(1 - \tilde{p});\quad n = \dfrac{1{,}64^2}{0{,}05^2} \cdot 0{,}6 \cdot 0{,}4 = 258$. Итак, n=258.

2. Для решения задачи используем замечание к формуле (3.19), из него следует, что при неизвестном $\tilde{p}$ пользуются оценкой: $\tilde{p}\tilde{q} = \tilde{p}(1 - \tilde{p}) \le 0{,}25$, тогда $n \le \dfrac{t^2}{4\varepsilon^2}$. В нашем случае $n \le \dfrac{1{,}64^2}{4 \cdot 0{,}05^2} \Rightarrow n \le 268{,}96$, следовательно n_{max}=269.

Ответ. 1. n=258; 2. n_{max}=269.

3.3. Задачи для самостоятельного решения

51.Коммерческий банк, изучая возможности предоставления долгосрочных кредитов населению, опрашивает своих клиентов для определения среднего размера такого кредита. Опрошены случайно выбранные 1000 клиентов. Среднее значение необходимого кредита в выборке составило 6750у.е. со стандартным отклонением 1460у.е. Найти границы доверительного интервала с надёжностью 0,95, для оценки неизвестного среднего значения кредита в генеральной совокупности.

52. На овцеводческой ферме из стада произведена выборка 36 овец для взвешивания. Их средний вес оказался равным 50кг. Предположив распределение веса нормальным и определив несмещённую оценку выборочной дисперсии $S^2=16$, найти доверительный интервал для оценки среднего веса во всём стаде с надёжностью: а)0,8; б)0,9; в)0.95. Сделайте выводы о зависимости величины интервала и надёжности оценки.

53. В течение года владельцем автостоянки было проведено 40 проверок. По данным проверок среднее число автомобилей, оставляемых на ночь на охрану, составило 400 единиц, а среднее квадратическое (стандартное) отклонение их числа 10 автомобилей. С вероятностью 0,99 оцените с помощью доверительного интервала истинное среднее число автомобилей, оставляемых на ночь на охрану. Обоснованы ли опасения владельца автостоянки, если по отчётности охранников среднее число автомобилей, оставляемых на ночь на охрану, составляет 395?

54. Среднее квадратическое отклонение нормально распределённой случайной величины X равно $\sigma(X)=1{,}5$, выборочная средняя $\overline{X}_в=12$, объём выборки *n*=49. Найдите доверительные интервалы для математического ожидания **M(X)** с заданной доверительной вероятностью γ, установите, как изменяется величина интервала в зависимости от величины надёжности:

$$а)\ \gamma=0{,}95; \quad б)\ \gamma=0{,}99; \quad в)\gamma=0{,}995.$$

55. Установите влияние объёма выборки на величину доверительного интервала для неизвестного математического ожидания нормально распределённой случайной величины X, если $\sigma(X)=6;\ \overline{X}_в=20{,}1;$ $\gamma=0{,}95$:

а) n=64; *б*) n=36;

в) n=144.

56. Служба контроля энергосбыта провела выборочную проверку расхода электроэнергии жителями одного из многоквартирных домов. Было отобрано 10 квартир и определён расход электроэнергии в течение одного из летних месяцев (квт.ч) 125, 78, 102, 140, 90, 45, 50, 125, 115, 112. С вероятностью 0,95 определите доверительный интервал для оценки среднего расхода электроэнергии на одну квартиру во всём доме.

57. Для определения потерь зерна в поле во время уборки урожая было взято 100 случайно отобранных проб. Выборочная средняя величина потерь составила 12,3*г/м*^{2}, среднее квадратическое отклонение составило 3,7*г/м*^{2}. С какой вероятностью можно утверждать, что ошибка в определении величины потерь зерна не превосходит 1*г* на квадратный метр?

58. Сколько дворов сельских жителей нужно обследовать в порядке случайного отбора из генеральной совокупности, чтобы определить среднюю урожайность картофеля с точностью до 5*ц/га*, с вероятностью 0,9973, если σ =20*ц/га.*

59. Выборочное исследование деятельности коммерческих банков некоторого региона показало, что в среднем каждый банк имеет 10 филиалов в регионе (со стандартным отклонением, равным 5). Найдите объём выборки, позволившей сделать такую оценку, если предельная ошибка выборочной средней находится в пределах 20% от её фактического значения, а доверительная вероятность составляет 0,95.

60. Для определения средней урожайности овса взято наудачу 20 проб по 1м^2 и определены $\overline{X}в = 0{,}125кг$ и $\sigma_в = 0{,}052кг$. Найти в каких границах заключена средняя урожайность с 1м^2 по всему полю, если вывод следует сделать с надёжностью 0,9.

61. Для изучения размера средней месячной зарплаты занятого населения региона произведена случайная выборка. Каким должен быть объём выборки, чтобы с доверительной вероятностью 0,997 можно было утверждать, что средняя месячная зарплата в выборке отличается от средней месячной зарплаты во всём регионе по абсолютной величине не более, чем на 25%, если средняя месячная заработная плата составляет 6200рублей со средним квадратическим отклонением 3000рублей.

62. В нескольких мелких магазинах проведена проверка качества 100изделий, после чего осуществлена обработка полученных данных. В результате получено несмещённое значение выборочной дисперсии S^2=16. Считая распределение качественных изделий нормальным, найти с надёжностью 0,95 доверительный интервал для оценки среднего квадратического отклонения.

63.Случайная величина *X* распределена по нормальному закону. Статистическое распределение выборки представлено в таблице:

x_i	3	5	7	8	10	12	14
m_i	3	7	4	6	7	5	8

Найти с надежностью 0,97 доверительный интервал для оценки математического ожидания и с надёжностью 0,95- для оценки среднего квадратического отклонения.

64.В выборке из 25 зёрен пшеницы средняя масса зерна составила $\overline{X}_в = 0{,}5г,\quad S^2 = 0{,}0025$(несмещённая оценка дисперсии). Предполагая, что случайная величина *X* (масса зерна) имеет нормальное распределение, найти с надёжностью 0,99: а) доверительный интервал для *M(X);*

б) доверительный интервал для неизвестного среднего квадратического отклонения;

в) доверительный интервал для *M(X),* если объём выборки равен 50. Сделайте практические выводы.

65. Максимальная толщина снежного покрова за последние 15 лет в данной местности по данным наблюдений была равна (в см.): 50; 48; 52; 53; 54; 61; 52; 50; 48; 54; 53; 50; 46; 53; 61. Найти доверительные интервалы для среднего значения толщины снежного покрова с надёжностью 0,95 и среднего квадратического отклонения с надёжностью 0,99, считая, что определяемая величина распределена по нормальному закону.

66. Выборочное обследование качества кирпича дало следующие результаты: из 1600проб в 32 случаях кирпич оказался бракованным. Требуется определить, в каких пределах заключается доля брака для всей продукции, если результат необходимо гарантировать с вероятностью 0,954.

67. При выборочном опросе 1200 телезрителей оказалось, что 456 из них регулярно смотрят программы телеканала НТВ. Построить 99%-й доверительный интервал, оценивающий долю всех телезрителей, предпочитающих программы НТВ.

68. Выборочное обследование малых предприятий города показало, что 95% малых предприятий в выборке относятся к негосударственной форме собственности. Принимая доверительную вероятность равной 0,954, определите в каких границах находится доля негосударственных малых предприятий в генеральной совокупности, если в выборку попали 100 предприятий.

69. Для определения процента вкладов не превышающих 10000рублей, произведена выборка 900 лицевых счетов. Среди них оказалось 30% вкладов не более 10000 каждый. С какой доверительной вероятностью можно утверждать, что процент таких вкладов в данном банке будет отличаться от найденного не более чем на 2% ?

70. Производится выборочное обследование доли лиц с высшим образованием в данной местности. Сколько нужно обследовать лиц, чтобы полученный результат гарантировать с вероятностью 0,95, при допустимой ошибке в определяемой доле 0,01.Решить эту же задачу, если ориентировочно известно, что процент лиц с высшим образованием равен примерно 4%.

71. Выборка объёмом 500 единиц произведена для определения процента всхожести зерна р. По выборке установлена относительная частота доброкачественных зёрен 0,94. Найти с какой вероятностью может принят в этом случае искомый процент всхожести, если допустимая погрешность в его определении равна $\pm 2\%$..

72. По результатам социологического обследования, при опросе 1500 респондентов рейтинг президента составил 30%.1). Найти границы, в которых с надёжностью 0,95 заключён рейтинг президента (при опросе всех жителей страны).2). Сколько респондентов надо опросить, чтобы с надёжностью 0,99 гарантировать предельную ошибку социологического обследования не более 1%.3). Тот же вопрос, если ни каких данных о рейтинге президента нет.

Тест 3

1.Если γ доверительная вероятность, то уровень значимости α равен:

а) $\frac{1-\gamma}{2}$; $\quad$ $б)1-\gamma$; $\quad$ $в)\frac{\gamma}{2}$; $\quad$ $г)1-\frac{\gamma}{2}$.

2.Установите соответствие между указанными доверительными интервалами и распределениями, использованными при их построении:

1).Доверительный интервал для $X_{ген.}$ при $n>30$	а) Распределение Стьюдента с (n –1) степенями свободы;
2).Доверительный интервал для дисперсии и среднего квадратического отклонения	б) нормальное распределение;
3).Доверительный интервал для $X_{ген.}$ с неизвестным σ и $n < 30$.	в) распределение χ^2 (хи-квадрат) с (n –1) степенями свободы.

3.Как изменяется величина доверительного интервала для $X_{ген}$ при увеличении объёма выборки?

а) увеличивается; $\quad$ б) уменьшается; $\quad$ в)остаётся неизменной.

4. При нахождении предельной ошибки выборки для оценки $X_{ген}$ при n>30 и доверительной вероятности γ =0,984, значение t, соответствующее этой вероятности равно….

5. Произведено 100 независимых испытаний, вероятность появления события A в каждом испытании одна и та же p. Событие A в этих испытаниях появилось 20 раз. Точечная оценка вероятности p , будет равна $\tilde{p} =$ ….

6. Для *задачи 5* данного теста найдите интервальную оценку для p с надежностью 0,9544:

а) (0,184; 0,216); б)(–0,08; 0,08); в)(0,12; 0,28).

7. С увеличением доверительной вероятности γ , доверительный интервал:

а) уменьшается; б) не изменяется; в)увеличивается.

8. Если доверительная вероятность равна γ при оценке дисперсии или среднего квадратического отклонения, то уровни значимости для χ_2^2 и χ_1^2 будут соответственно равны:

а) $\frac{1-\gamma}{2}; \frac{1+\gamma}{2}$, $б)1-\frac{1-\gamma}{2}; \frac{1-\gamma}{2}$, $в)\frac{1-\gamma}{2}; 1-\frac{1-\gamma}{2}$, $г)\frac{1+\gamma}{2}; \frac{1-\gamma}{2}$.

9. Как изменится объём выборки при построении доверительного интервала для $X_{ген.}$, если точность оценки повысить в 10 раз?

а) n увеличится в 10раз; б) n увеличится в 100раз;

в) n уменьшится в 10раз; г) n уменьшится в 100раз.

10. При нахождении предельной ошибки выборки для оценки $X_{ген.}$, n=10, доверительная вероятность $\gamma = 0{,}98$, тогда значение t соответствующее этим данным равно…

11. Произведена выборка из генеральной совокупности n=36. Найдены $\overline{X}_в$ =10 и S^2=9 (исправленная оценка дисперсии). Интервальная оценка для $X_{ген.}$ при доверительной вероятности 0,9544 задаётся границами:

а) (9;11), б) (7; 13), в) (9,5; 10,5).

12.Генеральная совокупность распределена нормально с $\sigma = 5$ Извлечена выборка, предельная ошибка выборки равна 0,5 при доверительной вероятности $\gamma = 0{,}9973$, тогда объём выборки *n* равен …

13.Из генеральной совокупности нормально распределённой *X*~*N*($a;\sigma$) сделана выборка. Какое распределение имеет средняя выборочная, вычисленная по выборке?

а)Стьюдента с (*n*–1) степенями свободы;

б) χ^2 (хи-квадрат) с (*n* –1) степенями свободы;

в) нормальное распределение;

г) стандартное нормальное распределение.

14. Если доверительный интервал для генеральной средней имеет вид: (3,3; 6,7), а $\overline{X}_в = 5$, то предельная ошибка выборки равна:

а) 3,4; б)1,7; в)5; г)2,5.

15.Какое распределение имеют статистики $\dfrac{nD_в}{\sigma^2} = \dfrac{(n-1)S^2}{\sigma^2}$?

а) нормальное,

б)Стьюдента с(*n* –1) степенями свободы,

в)стандартное нормальное,

г) χ^2 (хи-квадрат.) с (*n* –1) степенями свободы.

Вопросы для самопроверки по главе 3

1.В чём отличие точечной оценки числовой характеристики от интервальной?

2.Определение доверительного интервала.

3.Определение доверительной вероятности и уровня значимости.

4.Как определяется предельная ошибка выборки при нахождении доверительного интервала для генеральной средней?

5.Нахождение доверительного интервала для генеральной средней нормально распределённой генеральной совокупности при известном значении σ и неизвестном σ (n>30). Укажите какие таблицы Приложения 3 используются.

6.Построение доверительного интервала для $X_{ген.}$ по малой выборке (n<30).Укажите использованные таблицы Приложения 3.

7.Как связаны величина доверительного интервала и доверительная вероятность?

8.Зависимость ширины доверительного интервала от объёма выборки.

9.Нахождение объёма выборки по известной точности оценки для $X_{ген}$.

10.Доверительный интервал для генеральной дисперсии и среднего квадратического отклонения укажите два способа построения интервала. Какие таблицы Приложения 3 при этом используются?

11.Доверительный интервал для генеральной доли нормально распределённого количественного признака по выборочной доле.

12.Как построить доверительный интервал для генеральной доли по малой выборке?

Глава 4. Статистические гипотезы. Проверка статистических гипотез

4.1. Необходимый теоретический минимум

На разных этапах статистического исследования возникает необходимость в формулировании и экспериментальной проверке различных предположений о характеристиках конкретного массового явления.

Определение 4.1. ***Статистической гипотезой*** называют любое утверждение о виде неизвестного распределения или о параметрах известных распределений.

Примеры статистических гипотез:

1. генеральная совокупность распределена по показательному закону;
2. нормально распределенная случайная величина имеет среднее значение $a = 4$ (среднее квадратическое значение σ известно);
3. нормально распределенная случайная величина имеет среднее значение $a = 4$ (σ неизвестно);
4. нормально распределенная случайная величина имеет среднее значение $a \neq 4$ (σ известно);
5. дисперсии двух нормальных совокупностей равны между собой.

- В первой из вышеперечисленных гипотез выдвинуто предположение о виде неизвестного распределения; во второй, третьей и четвертой гипотезах – о параметре одного известного распределения; в пятой гипотезе – о параметрах двух известных распределений.

Определение 4.2. ***Проверкой статистических гипотез*** называется сопоставление высказанной гипотезы относительно генеральной совокупности с имеющимися выборочными данными, сопровождаемое количественной оценкой степени достоверности получаемого вывода.

Статистические гипотезы подразделяют на две группы:

I. *Нулевая (основная)* гипотеза – это выдвинутая гипотеза H_0 (т.е. проверяемое утверждение);

II. *Альтернативная (конкурирующая)* гипотеза – это гипотеза H_1, отрицающая или исключающая основную гипотезу.

- В примерах статистических гипотез, приведенных выше, вторая и четвертая гипотезы – конкурирующие. Коротко это можно записать так: $H_0 : a = 4$ (вторая гипотеза), $H_1 : a \neq 4$ (четвертая гипотеза).

Гипотезы о неизвестном параметре распределения бывают:

1. *Простые* – это гипотезы, в которых утверждается, что параметр θ имеет одно конкретное значение $(\theta = \theta_0)$.

2. *Сложные* – это гипотезы, в которых утверждается, что параметр θ имеет значение из совокупности значений $(\theta < \theta_0,\ \ \theta > \theta_0,\ \ \theta \neq \theta_0)$.

- В приведенных примерах статистических гипотез вторая гипотеза – простая, а третья и четвертая гипотезы – сложные.
- Гипотезу проверяют на основании выборки, полученной из генеральной совокупности. Из-за случайности выборки в результате проверки могут возникнуть ошибки и приниматься неправильные решения.

Виды ошибок, возникающих при проверке статистических гипотез:

1. *Ошибка первого рода* – имеет место тогда, когда отвергается правильная гипотеза H_0;
2. *Ошибка второго рода* – имеет место тогда, когда принимается неправильная гипотеза H_0.

- Вероятность ошибки первого рода обозначают α и называют *уровнем значимости* α, т.е. $\alpha = P(H_1 / H_0)$.
- Вероятность ошибки второго рода обозначают β, т.е. $\beta = P(H_0 / H_1)$.
- Вероятность принять верную гипотезу равна $P(H_0 / H_0) = 1 - \alpha$.

Вероятность отвергнуть неверную гипотезу H_0 (*мощность критерия*) равна $P(H_1 / H_1) = 1 - \beta$.

- Если α =0,01, то это означает, что в одном случае из 100 имеется риск допустить ошибку первого рода.

? Если уровень значимости α =0,05, то что это означает?

Определение 4.3. ***Статистическим критерием (статистикой)*** называется случайная величина *К*, служащая для проверки основной гипотезы H_0.

- Выбор критерия для проверки статистических гипотез может быть выполнен на основании различных принципов. Чаще всего для этого используют *принцип отношения правдоподобия,* который позволяет построить критерий, наиболее мощный среди всех возможных критериев.

Суть принципа правдоподобия: выбрать такой критерий (статистику) *К* с известной функцией плотности $f(k)$ при условии справедливости основной гипотезы H_0, чтобы при заданном уровне значимости α можно найти такую критическую точку $k_{кр}$ распределения $f(k)$, которая разделит область значений критерия на две области: *допустимую* (где результаты выборочного наблюдения выглядят наиболее правдоподобно) и *критическую* (где результаты выборочного наблюдения выглядят менее правдоподобно в отношении гипотезы H_0).

Определение 4.4. ***Наблюдаемым значением*** $K_{набл}$ называется значение статистики, вычисленное по выборке значений.

Множество значений определенной статистики разделяется на следующие два непересекающиеся подмножества (области):

1. *Допустимая область* (или *область принятия гипотезы*) – это множество значений статистики, при которых основная гипотеза H_0 принимается (не отклоняется).

2. *Критическая область* – это подмножество значений статистики, при которых основная гипотеза H_0 отвергается (отклоняется) и принимается альтернативная гипотеза H_1.

Определение 4.5. Точка $k_{кр}$, отделяющая критическую область от допустимой области, называется ***критической точкой.***

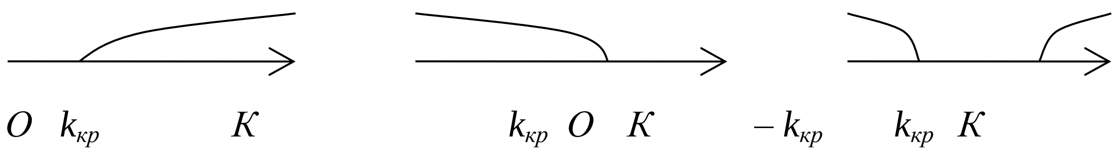
Виды критических областей

1. *Односторонняя* критическая область:

   а) *правосторонняя* – это критическая область, определяемая неравенством $k > k_{кр}$, где $k_{кр} > 0$ (рис. 4.1, а);

   б) *левосторонняя* – это критическая область, определяемая неравенством $k < k_{кр}$, где $k_{кр} < 0$ (рис. 4.1, б).

2. *Двусторонняя* критическая область – это область, определяемая двойным неравенством: $-k_{кр} < k < k_{кр}$ (рис. 4.1, в).



а) б) в)

Рис. 4.1

Алгоритм проверки статистической гипотезы:

1. Определить основную H_0 и альтернативную H_1 гипотезы.
2. Задать уровень значимости α.
3. Выбрать статистику.
4. Найти границы критической области (по альтернативной гипотезе H_1, по уровню значимости α – с помощью таблиц).
5. Вычислить наблюдаемое значение $K_{набл}$ статистики.
6. Сравнить значение статистики с критической областью.
7. Принять решение:

а) если значение статистики не принадлежит критической области, то гипотеза H_0 не отвергается;

б) если значение статистики принадлежит критической области, то гипотеза H_0 отвергается, а гипотеза H_1 принимается.

! Проверка каждого типа статистических гипотез выполняется с помощью соответствующего критерия, например:

– проверка гипотезы о *виде закона распределения* случайной величины осуществляется на основании *критерия согласия Пирсона* χ^2;

– проверка гипотезы о *равенстве неизвестных значений дисперсии* двух генеральных совокупностей – с помощью *критерия Фишера F*;

– проверка гипотез о *неизвестных значениях параметров* генеральных совокупностей – посредством *критерия Z нормально распределенной случайной величины* и *t –критерия Стьюдента*.

- В качестве примера использования статистического критерия приведем методику проверки гипотезы о равенстве среднего числовому значению (табл.4.1. и табл. 4.2)

Таблица 4.1

Проверка гипотезы о равенстве среднего числовому значению (дисперсия σ^2 генеральной совокупности известна)

Основная гипотеза H_0	Наблюдаемое значение критерия	Альтернативная гипотеза H_1	Условие для нахождения критических точек	Основная гипотеза H_0	
				не отвергается	отвергается
$H_0: a = a_0$	$K_{набл} = \dfrac{(\bar{x} - a_0)\sqrt{n}}{\sigma}$ где $\bar{x}$ – выборочное среднее; n – объем выборки	$H_1: a \neq a_0$	$\Phi(k_{кр}) = \dfrac{1-\alpha}{2}$	$\lvert k_{набл}\rvert < k_{кр}$	$\lvert k_{набл}\rvert > k_{кр}$
		$H_1: a > a_0$	$\Phi(k_{кр}) = \dfrac{1-2\alpha}{2}$	$k_{набл} < k_{кр}$	$k_{набл} > k_{кр}$

	($n \geq 30$)	$H_1 : a < a_0$	$\Phi(k'_{кр}) = \frac{1-2\alpha}{2}$ $k_{кр} = -k'_{кр}$	$k_{набл} > k_{кр}$	$k_{набл} < k_{кр}$

Таблица 4.2

Проверка гипотезы о равенстве среднего числовому значению (дисперсия σ^2 генеральной совокупности неизвестна)

Основная гипотеза H_0	Наблюдаемое значение критерия	Альтернативная гипотеза H_1	Условие для нахождения критических точек	Основная гипотеза H_0	
				не отвергается	отвергается
1	**2**	**3**	**4**	**5**	**6**
$H_0 : a = a_0$	$T_{набл} = \frac{(\bar{x} - a_0)\sqrt{n}}{s}$ где $\bar{x}$ - выборочное среднее, s – исправленное среднее квадратическое отклонение, n – объем выборки (n<30)	$H_1 : a \neq a_0$	$k_{кр} = t_{двуст}(\alpha; k)$ k=n–1	$\lvert T_{набл} \rvert < k_{кр}$	$\lvert T_{набл} \rvert > k_{кр}$
		$H_1 : a > a_0$	$k_{кр} = t_{прав}(\alpha; k)$ $k = n - 1$	$T_{набл} < k_{кр}$	$T_{набл} > k_{кр}$
		$H_1 : a < a_0$	$k'_{кр} = t_{прав}(\alpha; k)$ $k_{кр} = -k'_{кр}$	$T_{набл} > k_{кр}$	$T_{набл} < k_{кр}$

- Результаты проверки статистической гипотезы можно интерпретировать так: если приняли гипотезу H_1, то её можно считать доказанной а если приняли гипотезу H_0, то признали, что гипотеза H_0 не противоречит результатам наблюдений, т.е. в этом случае требуются дополнительные исследования.

4.2. Примеры решения типовых задач

Задача 4.1. Из нормальной генеральной совокупности с известным средним квадратическим отклонением $\sigma = 5{,}2$ извлечена выборка объема $n = 100$ и по ней найдена выборочная средняя $\bar{x} = 27{,}56$. Требуется при уровне значимости $\alpha = 0{,}05$ проверить основную гипотезу H_0: $a = 26$ при альтернативной гипотезе H_1: $a \neq 26$.

Решение. Методика проверки заданной основной гипотезы приведена в таблице 4.1:

1. $H_0 : a = 26 \quad (a_0 = 26)$.

2. $$K_{набл} = \frac{(\bar{x} - a_0)\sqrt{n}}{\sigma} = \frac{(27{,}56 - 26)\cdot\sqrt{100}}{5{,}2} = 3.$$

3. $H_1 : a \neq 26 \Rightarrow$ двусторонняя критическая область.

4. $$\Phi(k_{кр}) = \frac{1-\alpha}{2} = \frac{1 - 0{,}05}{2} = 0{,}475 \quad \Rightarrow \quad \Phi(k_{кр}) = 0{,}475.$$

Отсюда находим, что $k_{кр} = 1{,}96$ (нашли по таблице значений функции Лапласа (приложение). $\boxed{?}$

5-6. Сравним $|K_{набл}| = 3$ и $k_{кр} = 1{,}96$: так как $|k_{набл}| > k_{кр}$, то основная гипотеза H_0 отвергается.

Ответ. Выборочная и гипотетическая генеральная средние различаются значимо.

Задача 11.2. Составлена случайная выборка из 64 покупателей, которые интересовались товаром *А*. Из них этот товар купили 16 человек. Поставщик утверждает, что данный товар должен привлечь треть покупателей, а среднее квадратическое отклонение σ равно одному человеку. Проверить основную гипотезу при 5 % - м уровне значимости.

Решение. Предположим, что число покупателей, приобретающих товар *A*, есть случайная величина, подчиненная нормальному закону распределения. Из условия задачи следует, что предполагаемая генеральная средняя составит $a_0 = \frac{1}{3} \cdot 64 = 21$. Среднее квадратическое отклонение также известно $\sigma = 1$ (по условию). Итак, здесь следует проверить гипотезу о *равенстве среднего числовому значению при известной дисперсии.* (Так объем выборки $n = 64 > 30$ (табл. 4.1)) .

I. Запишем основную гипотезу: $H_0: a = 21$, то есть $a_0 = 21$.

II. Определим величину наблюдаемого значения статистики:

$$K_{набл} = \frac{(\bar{x} - a_0)\sqrt{n}}{\sigma} = \frac{(16 - 21)\sqrt{64}}{1} = -40.$$

III. Найдем границы критической области:

1. при альтернативной гипотезе $H_1: a \neq 21$ имеем *двустороннюю критическую область,* для которой

а) условием нахождения граничных точек $k_{кр}$ является:

$$\Phi(k_{кр}) = \frac{1 - \alpha}{2} = \frac{1 - 0{,}05}{2} = 0{,}475.$$

Откуда (по таблице значений функции Лапласа (*таблица* 3 *Приложения* 3) находим, что $k_{кр} = 1{,}96.$

- На рис. 4.2 построена данная критическая область.

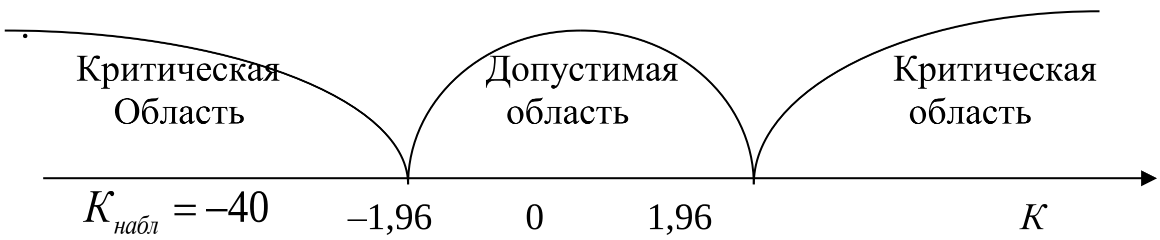


Рис. 4.2

б) Сравним $\left|K_{набл}\right|$ и $t_{кр}$. Получим $\left|K_{набл}\right| = 40 > k_{кр} = 1{,}96$. Значит, основная гипотеза $H_0: \ a = 21$ отвергается и принимается альтернативная гипотеза $H_1: \ a \neq 21$.

- На рис. 4.2 сделанный нами вывод о том, что гипотеза H_0 отвергается, проиллюстрирован наглядно. Действительно, наблюдаемое значение критерия $K_{набл} = -40$ принадлежит критической, а не допустимой области. А критическая область (согласно определению) – это та область, где принимается альтернативная гипотеза H_1.

2. При альтернативной гипотезе $H_1: \ a > 21$ *имеем правостороннюю критическую область* (рис. 4.3), для которой:

   а) условие для нахождения граничной точки:

$$\Phi\left(k_{кр}\right) = \frac{1-2\alpha}{2} = \frac{1-2\cdot 0{,}05}{2} = 0{,}45.$$

   По таблице значений функции находим, что $k_{кр} = 1{,}65.$

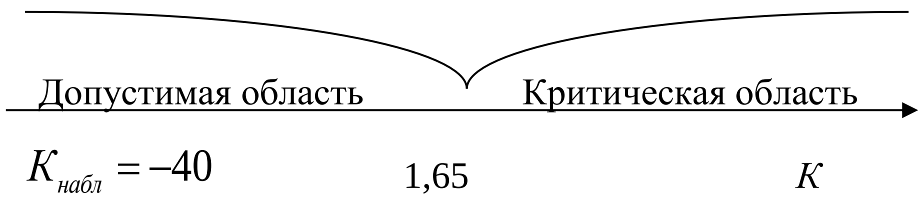


Рис.4.3

   б) Сравним $K_{набл}$ и $k_{кр}$. Получим, что $K_{набл} = -40 < k_{кр} = 1{,}65$.

   Значит, гипотеза H_0 не отвергается, а отвергается $H_1: \ a > 21$.

- На рис.4.3. наглядно видно, что наблюдаемое значение критерия не принадлежит критической области, т.е. альтернативная гипотеза отвергается.

? Если $K_{набл}$ принадлежит допустимой области, то это говорит о том, что основная гипотеза принимается безоговорочно или только о том, что она не отвергается (т.е. может быть принята, а может не быть принята)?

3. При альтернативной гипотезе вида $H_1:\ a < 21$ имеем *левостороннюю критическую область,* для которой:

а) найдем граничную точку критической области так:

– определим вспомогательную точку $k'_{кр}$ с помощью равенства

$\Phi(k'_{кр}) = \dfrac{1-2\alpha}{2}$, Откуда получаем, что $k'_{кр} = 1{,}65$;

– запишем собственно критическую точку:

$k_{кр} = -k'_{кр} = -1{,}65$ (геометрическая иллюстрация на рис.4.4).

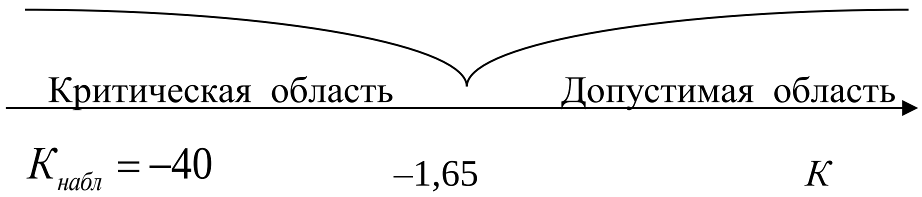


Рис.4.4

б) сравним $K_{набл}$ и $k_{кр}$: $K_{набл} = -40 < k_{кр} = -1{,}65$, т.е. гипотеза H_0 отвергается, принимается гипотеза $H_1:\ a < 21$.

- Действительно, наблюдаемое значение критерия $K_{набл}$ принадлежит критической, а не допустимой области (рис. 4.4).

! Обратите внимание, что при альтернативной гипотезе $H_1:\ a > 21$ основная гипотеза $H_0: a = 21$ не отвергается. На первый взгляд, данный вывод противоречит пункту 1.3. решения данной задачи, где доказано, что H_0 отвергается, но в действительности здесь противоречий нет. Так как, когда говорим «не отвергается», это не означает «принимается». В данном случае определенно можно сказать только, что альтернативная гипотеза $H_1:\ a > 21$ неверна. Таким образом, уточняется гипотеза $H_0: a \neq 21$ ($a > 21$ или $a < 21$).

Ответ. Утверждение от поставщика о том, что товар привлечет треть покупателей, неверно. В действительности (при 5%-м уровне значимости) товар привлечет внимание менее трети покупателей.

- При 1%-м уровне значимости ($\alpha = 0{,}01$) и при известной дисперсии генеральной совокупности границы критической области находятся так:

а) для двусторонней критической области:

$$\Phi(k_{кр}) = \frac{1-0{,}01}{2} = 0{,}495, \qquad \text{т.е.} \qquad k_{кр} = 2{,}58;$$

б) для правосторонней критической области:

$$\Phi(k_{кр}) = \frac{1-2\cdot 0{,}01}{2} = 0{,}49, \qquad \text{т.е.} \qquad k_{кр} = 2{,}33;$$

в) для левосторонней критической области: $k_{кр} = -2{,}33$.

Задача 4.3. По выборке объема $n = 16$, извлеченной из нормальной генеральной совокупности, найдены выборочная средняя $\bar{x} = 118{,}2$ и «исправленное» среднее квадратическое отклонение $s = 3{,}6$. При уровне значимости $\alpha = 0{,}05$ проверить основную гипотезу H_0: $a = 120$ при альтернативной гипотезе $H_1:\ a \neq 120$.

Решение. Методика проверки заданной основной гипотезы приведена в таблице 4.2:

1. $H_0: a = 120$ ($a_0 = 120$).

2. $T_{набл} = \dfrac{(\bar{x} - a_0)\sqrt{n}}{s} = \dfrac{(118{,}2 - 120)\cdot\sqrt{16}}{3{,}6} = -2.$

3. H_1: $a \neq 120$, т. е. критическая область двусторонняя.

4. $k_{кр} = t_{двуст}(\alpha; k) = t_{двуст}(0{,}05;15) = 2{,}13$ – нашли по таблице критических точек распределения Стьюдента для двусторонней критической области (*Таблица* 4 *Приложения* 2).

5 – 6. Сравним $|T_{набл}| = 2$ и $k_{кр} = 2{,}13$: так как $|T_{набл}| < k_{кр}$, то основная гипотеза не отвергается.

Ответ. Выборочная средняя $\bar{x} = 118{,}2$ от гипотетической генеральной средней отличается незначимо.

Задача 4.4. Проектный контролируемый размер изделий, изготавливаемых станком–автоматом $a_0 = 35 мм$. Измерения 20 случайно отобранных изделий дали следующие результаты:

Контролируемый размер x_i	34,8	34,9	35,0	35,1	35,3
Число изделий m_i	2	3	4	6	5

При 5%-м уровне значимости проверить гипотезу $H_0 : a = a_0$ при альтернативной гипотезе H_1: $a \neq a_0$.

Решение.

I. Основная гипотеза имеет вид: $H_0 : a = 35$, альтернативная – $H_1 : a \neq 35$.

II. В данном случае дисперсия генеральной совокупности неизвестна (объем выборки $n = 20 < 30$), поэтому в качестве критерия применяем t – критерий Стьюдента, поэтому наблюдаемое значение статистики вычисляем по формуле: $T_{набл} = \dfrac{(\bar{x} - a_0)\sqrt{n}}{s}$, где n - объем выборки; $\bar{x}$ - выборочное среднее; s – исправленное среднее квадратическое отклонение.

1) Найдем средний размер изделий выборки, т.е. выборочное среднее $\overline{x}$:

$\overline{x} = \frac{701{,}4}{20} = 35{,}07$. (Сумма $\sum_{i=1}^{5} x_i m_i$ найдена в таблице 4.2)

Таблица 4.2

i	x_i	m_i	$x_i m_i$	$x_i^2 m_i$
1	34,8	2	?	?
2	34,9	3	?	?
3	35,0	4	?	?
4	35,1	6	?	?
5	35,3	5	?	?
$\sum$	-	20	701,4	24598,62

2) Найдем «исправленную» дисперсию по формуле (2.21):

$s^2 = \frac{n}{n-1} \cdot D_в$, где $D_в$ – выборочная дисперсия (взвешенная).

По формуле 2.13 получаем:

$$D_в = \sum_{i=1}^{5} x_i^2 m_i - \overline{x}^2 = \frac{1}{2} \cdot 24598{,}62 - (35{,}07)^2 = 1229{,}931 - 1229{,}9049 = 0{,}0261.$$

Таким образом, получаем, что $s^2 = \frac{20}{19} \cdot 0{,}0261 \approx 0{,}027.$

3) «Исправленное» среднее квадратическое отклонение равно:

$s = \sqrt{0{,}027} \approx 0{,}16$.

4) Итак, наблюдаемое значение статистики равно:

$$T_{набл} = \frac{(35{,}07 - 35{,}0)\sqrt{20}}{0{,}16} = 1{,}96.$$

III. Границы *двусторонней критической области* найдем по таблице критических точек распределения Стьюдента (*Таблица* 6 *Приложения* 2).

По уровню значимости $\alpha = 0{,}05$, помещенному в верхней строке вышеозначенной таблицы, и по числу степеней свободы $k = n - 1 = 20 - 1 = 19$ находим критическую точку $k_{кр} = t_{двуст.кр}(0{,}05;19) = 2{,}09$.

Так как $|T_{набл}| = 1{,}96 < k_{кр} = 2{,}09$ (рис.4.5), то оснований отвергать нулевую гипотезу нет.

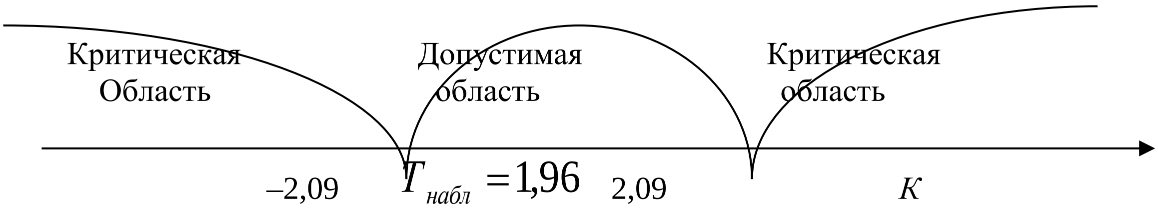


Рис. 4.5

Ответ. Станок обеспечивает проектный размер изделий.

- Еще раз акцентируем внимание на том, что если основную гипотезу H_0 нельзя отклонить, то это не означает, что высказанное предположение о генеральной совокупности является единственно подходящим: просто ему не противоречат имеющиеся выборочные данные, но таким же свойством наряду с высказанной гипотезой могут обладать и другие статистические гипотезы.

Задача 4.5. Техническая норма предусматривает в среднем 40 секунд на выполнение определенной технологической операции на конвейере по производству часов. От работающих на этой операции поступили жалобы, что они в действительности затрачивают на неё больше времени. Для проверки данной жалобы произведены хронометрические измерения времени выполнения этой технологической операции у 16 работниц, занятых на ней, и получено среднее время выполнения операции $\bar{x} = 42$ (секунд). Можно ли по имеющимся хронометрическим данным на уровне значимости $\alpha = 0{,}01$ отклонить гипотезу о том, что среднее время выполнения этой операции соответствует норме, если:

1) «исправленное» выборочное среднее квадратическое отклонение составляет $s = 3{,}5$ (секунд);

2) выборочное среднее квадратическое отклонение равно $\sigma_в = 3{,}5$ (секунд)?

Решение.

1) Здесь необходимо проверить гипотезу о том, что неизвестная генеральная средняя нормальной совокупности точно равна определенному числу, когда дисперсия генеральной совокупности неизвестна (объем выборки $n = 16 < 30$).

а) *Основная гипотеза* $H_0 : a = 40$ ($a_0 = 40$), т.е. неизвестное математическое ожидание a равно предполагаемому числовому значению a_0 (применительно к условию данной задачи – *время выполнения технологической операции соответствует норме*).

Альтернативная гипотеза $H_1 : a > 40$, т.е. неизвестное математическое ожидание a больше числового значения a_0 (применительно к условию данной задачи – *время выполнения технологической операции больше установленной нормы*).

б) Критерием для проверки гипотезы в данном случае является t – *критерий Стьюдента.* Определим наблюдаемое значение данного критерия:

$$T_{набл} = \frac{(\overline{x} - a_0)\sqrt{n}}{s} = \frac{42 - 40}{3{,}5} \approx 2{,}286.$$

в) Найдем по таблице критических точек распределения Стьюдента (*Таблица* 6 *Приложения* 2) критическую точку *правосторонней критической области* по уровню значимости $\alpha = 0{,}01$ и числу степеней свободы . $k = n - 1 = 16 - 1 = 15$. Получим следующее:

$$k_{кр} = t_{прав}(\alpha; k) = t_{прав}(0{,}01;15) = 2{,}6$$

г) Сравним $T_{набл}$ и $k_{кр}$: $T_{набл} = 2{,}286 < k_{кр} = 2{,}6$, т.е. на заданном уровне значимости основная гипотеза не отклоняется.

- В условиях данной задачи полученный вывод можно сформулировать так: по имеющимся хронометрическим данным на уровне значимости $\alpha = 0{,}01$ нельзя отклонить гипотезу о том, что среднее время выполнения этой операции соответствует норме. Значит, жалобы работниц необоснованны.

2) В этом случае также следует проверить гипотезу о том, что неизвестная генеральная средняя нормальной совокупности точно равна определенному числу (при неизвестной дисперсии генеральной совокупности). Алгоритм решения задачи будет тот же, что и в первом случае. Отличие будет в формуле для расчета наблюдаемого значения критерия, так как здесь задано выборочное среднее квадратическое отклонение $\sigma_в$ (а не «исправленное» *s*). По формуле (2.21) имеем:

$$s^2 = \frac{n}{n-1} \cdot D_в \quad \Rightarrow \quad \frac{n}{s^2} = \frac{n-1}{\sigma_в{}^2}.$$

Извлечем из обеих частей последнего равенства квадратный корень:

$$\frac{\sqrt{n}}{s} = \frac{\sqrt{n-1}}{\sigma_в}.$$

Тогда получаем:

$$T_{набл} = \frac{(\bar{x} - a_0)\sqrt{n}}{s} = \frac{(\bar{x} - a_0) \cdot \sqrt{n-1}}{\sigma_в} = \frac{42-40}{3{,}5}\sqrt{15} \approx 2{,}213.$$

В пункте (в) части 1) решения задачи нашли, что $k_{кр} = 2{,}6$.

Сравним величины $T_{набл}$ и $k_{кр}$.

Получаем, что, $T_{набл} = 2{,}213 < k_{кр} = 2{,}6$,

значит, основная гипотеза H_0 не отвергается.

Ответ.

1), 2) Гипотеза о том, что среднее время выполнения технологической операции соответствует норме, не отклоняется. Следовательно, жалобы работниц на то, что нормы выполнения рассматриваемой технологической операции завышены, необоснованны.

4.3 Задачи для самостоятельного решения

73. Для нормальной генеральной совокупности с известным средним квадратическим отклонением $\sigma = 6{,}24$ извлечена выборка объема $n = 100$ и по ней найдена выборочная средняя $\bar{x} = 68{,}21$. Требуется при уровне значимости $\alpha = 0{,}01$ проверить основную гипотезу $H_0 : a = 67$ при альтернативной $H_1 : a \neq 26$.

74. По выборке объема $n = 9$, извлеченной из нормальной генеральной совокупности найдены выборочная средняя $\bar{x} = 126{,}5$ и «исправленная» дисперсия $S^2 = 2{,}25$. При уровне значимости $\alpha = 0{,}05$ проверить основную гипотезу $H_0 : a = 125{,}5$ при альтернативной гипотезе $H_1 : a \neq 125{,}5$.

75. Фирма-поставщик в рекламном буклете утверждает, что средний срок безотказной работы предлагаемого изделия – 2900 ч. Для выборки из 50 изделий средний срок безотказной работы оказался равным 2720 ч при «исправленном» среднем квадратичном отклонении 700 ч. При 5%-м уровне значимости проверить гипотезу о том, что значение 2900 ч является математическим ожиданием.

76. По результатам 10 замеров установлено, что среднее время обслуживания мастером клиента $\bar{x} = 15$ *мин.* Предполагая, что время обслуживания клиента – нормально распределенная случайная величина с дисперсией $\sigma_x^2 = 9\ мин^2$, при уровне значимости $\sigma = 0{,}05$

установить, можно ли принять в качестве норматива (математического ожидания) для обслуживания одного клиента: а) 21 мин; б) 16 мин.

77. В таблице приведена динамика среднесуточных привесов крупного рогатого скота (в граммах) в период 2000-2004 гг. (на начало года)[1] в сельскохозяйственных организациях Приморского края:

Год	2000	2001	2002	2003	2004
Среднесуточный привес КРС (граммов)	168	113	265	269	274

В средствах массовой информации было заявлено, что средний привес КРС в период 2000-2010 гг. составит 220 граммов.

1) На основе имеющейся информации при уровне значимости $\alpha = 0{,}01$ проверьте истинность этого утверждения.

2) Изменится ли полученный в пункте 1 вывод, если уровень значимости принять 0,05. Объясните результаты.

---

78. Из нормальной генеральной совокупности с известным средним квадратическим отклонением $\sigma = 3{,}21$ извлечена выборка объема $n = 121$ и по ней найдена выборочная средняя $\bar{x} = 32{,}5$. Проверить (при уровне значимости $\alpha = 0{,}05$) основную гипотезу $H_0 : a = 33$ при альтернативной $H_1 : a \neq 33$.

79. По выборке объема $\mathrm{n} = 16$, извлеченной из нормальной генеральной совокупности, найдены выборочная средняя $\bar{x} = 91{,}43$ и «исправленное» среднее квадратическое отклонение $s = 1{,}26$. При уровне значимости $\alpha = 0{,}01$ проверить основную гипотезу $H_0 : a = 90$ при альтернативной гипотезе $H_1 : a \neq 90$.

[1] По данным Приморского краевого комитета государственной статистики

80. Поставщик удобрений утверждает, что применение новой партии удобрений обеспечивает урожайность пшеницы в 60 ц/га. Удобрения внесли на площади в 37 га и получили урожай 55 ц/га при «исправленном» среднем квадратичном отклонении 3 ц/га. При 5 %-м уровне значимости оценить справедливость утверждения поставщика.

81. В таблице приведена динамика наличия техники в сельскохозяйственных организациях Приморского края[2]:

	2000	2001	2002	2003	2004
Тракторы (без тракторов, на которых смонтированы землеройные, мелиоративные и другие машины)	31	23	25	36	33
Комбайны зерноуборочные	6	5	5	7	7
Комбайны кормоуборочные	4	2	2	2	2

При уровне значимости $\alpha = 0{,}01$ проверить истинность следующих утверждений:

1) среднее число тракторов в сельхозпредприятиях Приморского края за последние десять лет равно 34 единицам;

2)комбайнов (зерноуборочных и кормоуборочных) 6 единиц.

82. Компания, занимающаяся консультированием в области инвестиций, заявляет, что среднегодовой процент по акциям определенной отрасли промышленности составляет 11,5%. Инвестор, желая проверить истинность этого утверждения, на основе случайной выборки 50 акций выявил, что среднегодовой процент по ним составил 10,8% с исправленным средним квад-

ратическим отклонением s=3,4%. На основе имеющейся информации определите, имеет ли инвестор достаточно оснований, чтобы опровергнуть заявление компании? Принять уровень значимости α =0,05.

[2] По данным Приморского краевого комитета государственной статистики

83. Средний диаметр подшипников должен составлять 35мм. Однако для выборки из 82 подшипников он составил 35,3 мм при «исправленном» среднем квадратичном отклонении 0,1мм. При 5%-м уровне значимости проверить гипотезу о том, что станок, на котором изготавливаются подшипники, не требует подналадки.

84. Компания, производящая средства для потери веса, утверждает, что прием таблеток в сочетании со специальной диетой позволяет сбросить в среднем в неделю 400г веса. Отобрано 25 человек, использующих эту терапию, и обнаружено, что в среднем еженедельная потеря в весе составила 430 г со средним квадратическим отклонением 110г. Проверьте гипотезу о том, что средняя потеря в весе составляет 400г ($\alpha = 0{,}05$).

85. Среднесуточная продажа хлеба в течение многих лет для магазина составила 6 т при среднем квадратичном отклонении 0,05 т. Сегодня магазином было продано 7 т хлеба. Можно ли при 5%- м уровне значимости предполагать, что завтра будет продано 7 т хлеба?

86. Фирма - изготовитель женских украшений, выпустив новый товар, утверждает, что 40 % покупателей купят эти украшения. В ходе 10-дневной рекламной распродажи в среднем приобрели украшения 29,5 % покупателей, «исправленное» среднее квадратичное отклонение составило 16,5 %. При 5% - м уровне значимости оценить утверждение изготовителя товара.

87. Компания, выпускающая в продажу новый сорт растворимого кофе, провела проверку вкусов покупателей по случайной выборке из 400 человек и выяснила, что 220 из них предпочли новый сорт всем остальным. Проверить при $\alpha = 0{,}01$ гипотезу о том, что, по крайней мере, 52% потребителей предпочтут новый сорт кофе.

88. Страховая компания изучает вероятность дорожных происшествий для подростков, имеющих мотоциклы. За прошедший год проведена случайная выборка 2000 страховых полисов подростков - мотоциклистов и выявлено, что 15 из них попали в дорожные происшествия и предъявили компании требование о компенсации за ущерб. Можно ли отклонить гипотезу о том, что менее 1 % всех подростков-мотоциклистов, имеющих страховые полисы, попадали в дорожные происшествия в прошлом году? (При $\alpha = 0{,}05$).

89. Из большой партии ананасов одного размера случайным образом отобрано 36 штук. Выборочная средняя масса оной штуки при этом оказалась равной 930г. Используя двусторонний критерий при $\alpha = 0{,}05$, проверить гипотезу, что средняя масса одного ананаса (по утверждению поставщика) составляет 1 кг, если: а) среднее квадратичное отклонение известно и составляет 200г; б) среднее квадратичное отклонение неизвестно, а «исправленное» составило 250г.

90. Компания по производству безалкогольных напитков предполагает выпустить на рынок новую модификацию популярного напитка, в котором сахар заменен сукразитом. Компания хотела бы быть уверенной в том, что не менее 70% ее потребителей предпочтут новую модификацию напитка. Новый напиток был предложен на пробу 2 тыс. чел., и 1422 из них сказали, что он вкуснее старого. Может ли компания отклонить предложение о том, что только 70% всех ее потребителей предпочтут новую модификацию напитка старой? Принять уровень значимости $\alpha = 0{,}05$.

91. Производители нового типа аспирина утверждают, что он снимает головную боль за 30 мин. Случайная сборка 100 чел., страдающих головными болями, показала, что новый тип аспирина снимает головную боль за 28,6 мин при среднем квадратическом отклонении 4,2 мин. Проверьте на уровне значимости $\alpha = 0{,}05$ справедливость утверждения производителей аспирина о том, что это лекарство излечивает головную боль за 30 мин.

92. Проверить (при уровне значимости 1%) утверждение о том, что средняя дебиторская задолженность сельскохозяйственных организаций

Приморского края в период 2000-2010 гг. составит 4 млн руб. если известны следующие фактические данные:

	2000	2001	2002	2003	2004
Дебиторская задолженность (тыс. руб.)	4291	3436	8783	2515	2348

93. Динамика внесения минеральных удобрений в сельскохозяйственных организациях Приморского края приведена в таблице:

Год	2000	2001	2002	2003	2004
Внесено минеральных удобрений на 1 гектар посева, кг	26,0	34,5	9,2	17,4	13,6

Проверить утверждение о том, что в период 2000-2010 гг. в среднем в год внесено 22 кг минеральных удобрений ($\alpha = 0{,}05$).

Тест 4

1. Статистической гипотезой называют:

а) любое утверждение о физических свойствах объекта;

б) некоторое утверждение о виде неизвестного распределения или о параметрах известного распределения;

в) любое утверждение о виде известного распределения или о параметрах неизвестного распределения;

г) утверждение о виде распределения.

2. Ошибка первого рода имеет место в том случае, если:

а) отвергается правильная основная гипотеза H_0;

б) принимается неправильная основная гипотеза H_0;

в) отвергается правильная альтернативная гипотеза H_1;

г) принимается неправильная альтернативная гипотеза H_1.

3. Подмножество значений статистики, при которых основная гипотеза H_0 не отклоняется называется:

а) областью вероятных значений; б) областью значений статистики;

в) критической областью; г) допустимой областью.

4. Проверяется гипотеза о равенстве среднего числовому значению $H_0 : a = a_0$ (объем выборки $n = 25$). Тогда применяется следующая статистика:

а) $T = \frac{(\overline{x} - a_0)\cdot\sqrt{n}}{s}$ (где S – исправленное среднее квадратическое отклонение, $\overline{x}$ – выборочное среднее),

б) $K = \frac{(\overline{x} - a_0)\cdot\sqrt{n}}{\sigma}$ (где σ – выборочное среднее квадратическое отклонение; $\overline{x}$ – выборочное среднее),

в) $T = \frac{(\overline{x} - a_0)\cdot S}{\sqrt{n}}$ (где S – исправленное среднее квадратическое отклонение, $\overline{x}$ – выборочное среднее);

г) $T = \frac{(\overline{x} - a_0)\cdot\sigma}{\sqrt{n}}$ (где σ – выборочное среднее квадратическое отклонение; $\overline{x}$ – выборочное среднее).

5. Из нормальной генеральной совокупности со средним квадратическим отклонением $\sigma = 5{,}4$ извлечена выборка объема $n = 121$ и по ней найдена выборочная средняя $\overline{x} = 37{,}22$. Проверяется гипотеза $H_0 : a = 36$. Тогда наблюдаемое значение критерия (с точностью до двух знаков после запятой) равно…

6. Проверяется гипотеза $H_0 : a = a_0$ при альтернативной гипотезе $H_1 : a \neq a_0$. Наблюдаемое значение критерия $T_{набл.} = -4$, а критическая точка (при заданном уровне значимости) составляет $k_{кр} = 2{,}13$. Тогда основная гипотеза H_0 :

а) отвергается; б) не отвергается;

в) нужны таблицы распределения Стьюдента;

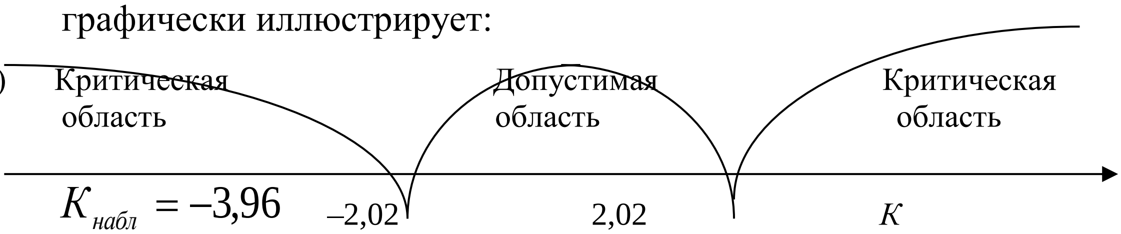
г) нужны дополнительные опыты.

7. Проверяется гипотеза H_0 : a = a_0 (объем выборки n = 100). Установить соответствие между видом альтернативной гипотезы и условием для нахождения граничных точек критической области (при заданном уровне значимости α).

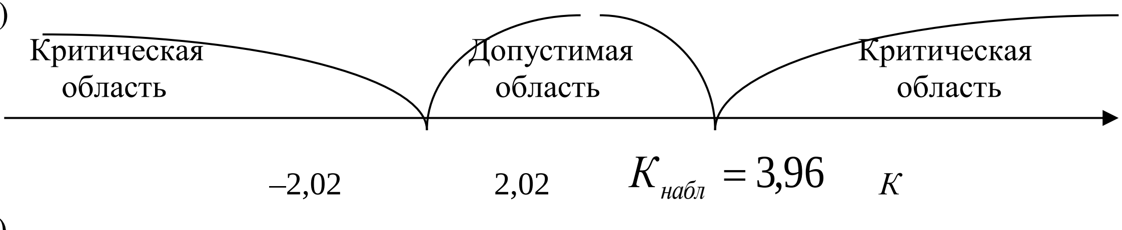
Вид альтернативной гипотезы	Условие для нахождения критических точек $k_{\text{кр}}$
1. H_1: $a \neq a_0$	а) $\Phi\left(k_{кр}\right) = \dfrac{1-2\alpha}{2}$, где $\Phi(x)$ –функция Лапласа;
2. H_1 : $a > a_0$	б) $k_{\text{кр}} = -k'_{\text{кр}}$, где $\Phi\left(k'_{кр}\right) = \dfrac{1-2\alpha}{2}$;
3. H_1 : $a < a_0$	в) Φ ($k_{кр}$) = $\dfrac{1-\alpha}{2}$.

8. Основная гипотеза $H_0 : a = a_0$, при альтернативной гипотезе H_1: $a < a_0$ отвергается (дисперсия генеральной совокупности известна), что графически иллюстрирует:

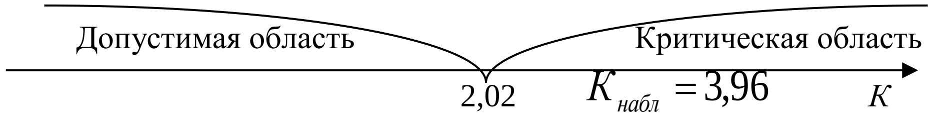
а)



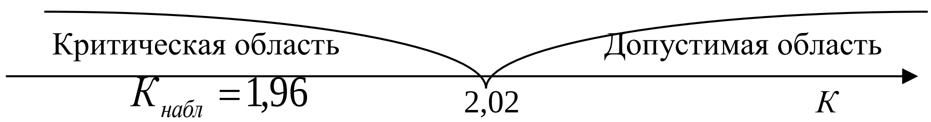
б)



в)



г)



9. Гипотеза, которая утверждает, что параметр θ имеет значение из совокупности значений, называется:

а) простой; б) нулевой; в) сложной; г) альтернативной.

10. Из нормальной генеральной совокупности извлечена выборка объемом $n=17$. Проверяется основная гипотеза $H_0 : a = 45$. Найти наблюдаемое значение критерия T, если выборочная дисперсия $D_в = 4$, а выборочное среднее $\overline{x_в} = 47$.

Вопросы для самопроверки по главе 4

1. Понятие и виды статистических гипотез.

2. Виды ошибок, возникающие при проверке гипотез.

3. Уровень значимости; мощность критерия.

4. Понятие статистического критерия, наблюдаемое значение критерия.

5. Допустимая и критическая области, критическая точка.

6. Общий вид алгоритма проверки статистических гипотез.

7. Методика проверки гипотезы о равенстве среднего числовому значению при:

а) известной дисперсии генеральной совокупности;

б) неизвестной дисперсии генеральной совокупности.

Глава 5. Элементы корреляционного анализа

5.1. Необходимый теоретический минимум

Функциональная, статистическая, корреляционная зависимости

Функциональная зависимость между двумя переменными характеризуется тем, что каждому значению одной переменной соответствует вполне определённое значение другой переменной.

- *Примеры функциональных зависимостей:* путь и время; количество купленного товара и его стоимость; количество потреблённой энергии и плата за неё.

- В экономике, в большинстве случаев, между переменными существует зависимость, когда каждому значению одной переменной соответствует несколько значений другой переменной, встречающихся не одинаково часто. Иными словами, каждому значению одной переменной соответствует распределение другой переменной, то есть рассматривается связь между случайными величинами. Строгая функциональная зависимость реализуется редко, так как обе величины или одна из них подвержена ещё действию случайных факторов.

Определение 5.1. Зависимость между случайными величинами X и Y, состоящая в том, что каждому значению одной величины соответствует распределение другой величины, называется ***статистической зависимостью***.

- К переменным величинам, связанным статистической зависимостью относятся, например: урожайность и количество внесённых удобрений, рост человека и его масса, объём дневной продукции и её себестоимость по различным предприятиям. На практике часто используют связь между изменениями одной случайной величины X и изменениями математического ожидания (средней) другой величины Y, то есть $M(Y/X=x)=f(x)$ или $M(X/Y=y)=\varphi(y)$.

Определение 5.2. Зависимость между двумя переменными величинами состоящая в том, что при изменении одной величины изменяется среднее значение (математическое ожидание) другой величины называется ***корреляционной звисимостью***.

В математической статистике имеем дело с оценками числовых характеристик, поэтому в качестве оценки условного математического ожидания принимают условное среднее***. Условным средним*** $\overline{y}_x$ называется среднее арифметическое наблюдаемых значений *Y*, соответствующих *X*=*x*. *Например*, если при *X*=3, величина *Y* принимает значения $y_1=2$, $y_2=7$, $y_3=6$, то $\overline{y}_x = \frac{2+7+6}{3} = 5$. Условные средние являются функциями соответственно от x и y.

Выборочным уравнением регрессии Y на X **называется зависимость вида:**

$$\overline{y}_x = f^*(x). \tag{5.1}$$

Соответственно, *выборочным уравнением регрессии X на Y* **называется зависимость вида**

$$\overline{x}_y = \varphi^*(y). \tag{5.2}$$

Графики соответствующих функций $f^*(x)$, $\varphi^*(y)$ называются ***выборочными линиями регрессии.***

Статистическую связь между переменными можно изучать методами корреляционного и регрессионного анализа. Основной задачей регрессионного анализа является установление формы зависимости между переменными и изучение этой зависимости.

Основная задача корреляционного анализа – выявление связи между переменными и оценка её тесноты. Для решения этих задач используют соответствующий математический аппарат. Рассмотрим задачу в общем виде.

Пусть имеется два ряда наблюдаемых зависимых между собой величин *X* и *Y*. Если x_i и y_i встречаются по одному разу, то их значения записывают в виде таблицы 5.1.

Таблица 5.1

X_i	x_1	x_2	x_3	…	x_k	…	x_n
Y_i	y_1	y_2	y_3	…	y_k	…	y_n

Если каждому значению y_i отвечает несколько значений *x*, а каждому x_j – несколько значений *y*, то эти данные записывают в виде таблицы 5.2 .

Таблица 5.2

Y / X	x_1	x_2	…	x_k	$\sum\limits_{j=1}^{k}$
y_1	n_{11}	n_{12}	…	n_{1k}	$\sum n_{1j} = l_1$
y_2	n_{21}	n_{22}	…	n_{2k}	$\sum n_{2j} = l_2$
…	…	…	…	…	…
y_m	n_{m1}	n_{m2}	…	n_{mk}	$\sum n_{mj} = l_m$
$\sum\limits_{i=1}^{m}$	$\sum n_{i1} = n_1$	$\sum n_{i2} = n_2$	…	$\sum n_{ik} = n_k$	$n = \sum\limits_{i=1}^{m} l_i = \sum\limits_{j=1}^{k} n_j$

- В таблице 5.2. n_{ij} – частота, показывающая число повторений парные значений x_i и y_j . Если парные значения не появляются ни разу, то в соответствующей клетке ставят прочерк.

Таблица 5.2, в которой результаты наблюдений записаны в порядке возрастания с указанием частоты n_{ij}, называется ***корреляционной таблицей.***

Корреляционная таблица может быть составлена как для дискретных так и для непрерывных признаков. В последнем случае строят интервальный ряд для значений *X, Y* затем переходят к дискретному ряду, заменяя соответствующий интервал $(x_{i-1};x_i)$ его серединой $x_i^* = \dfrac{x_{i-1} + x_i}{2}$. Также поступают со случайной величиной *Y*.

Графически, если взаимосвязь двух признаков задана таблицей 5.1, то в системе координат строят точки ($x_i;y_i$) и по расположению точек судят о форме зависимости (линейная, показательная, параболическая и т.д.). Чем сильнее связь между признаками, тем теснее будут группироваться точки вокруг определённой линии, выражающей форму связи (см. рис. 5.1).

- Данные корреляционной таблицы также можно изобразить графически. Для этого каждую пару значений ($x_i;y_j$) изображают точкой в системе координат; частоту n_{ij}, с которой эта пара встречается в таблице, изображают в виде числа, помещённого возле соответствующей точки.
- Графическое изображение корреляционной таблицы называется ***полем корреляции***. На таком графике наглядно можно проследить общую тенденцию в изменении Y с изменением X (см. рис. 5.2).

Отыскание параметров выборочного уравнения линейной регрессии

Пусть известны результаты опыта, целью которого является исследование зависимости между двумя величинами X и Y (например, величина прибыли и объём инвестиций, изменение по месяцам курса доллара и т. д.). Предположим, что различные значения x признака X и соответствующие значения y признака Y наблюдаются по одному разу (см. табл. 5.1).Строим точки ($x_i;y_i$) на координатной плоскости.

Предположим, что точки расположены приближённо вдоль некоторой прямой, уравнение которой ищем в виде:

$$y=a+bx\ . \qquad \textbf{(5.3)}$$

Среди всех прямых линий $y=a+bx$ ищут наиболее близкую к данной системе точек, причём близость будем измерять суммой квадратов отклонений:

$$S=\sum_{i=1}^{n}\left[y_i-\left(a+bx_i\right)\right]^2 .$$

Из всех прямых выбирают ту, для которой сумма сумма квадратов отклонений S минимальна.

Так как минимизируется сумма квадратов отклонений экспериментальных и теоретических значений, то предложенный метод называют *методом наименьших квадратов*. А прямую построенную по этому методу, называют *МНК-прямой*. Для определения параметров уравнения прямой линии y на x , получают систему *нормальных уравнений*.

Система нормальных уравнений для отыскания неизвестных параметров прямой линии регрессии имеет вид:

$$\begin{cases} \left(\sum_{i=1}^{n} x_i^{\ 2}\right)\cdot b + \left(\sum_{i=1}^{n} x_i\right)\cdot a = \sum_{i=1}^{n} x_i \cdot y_i \\ \left(\sum_{i=1}^{n} x_i\right)\cdot b + na = \sum_{i=1}^{n} y_i \end{cases} . \qquad \textbf{(5.4)}$$

Решая систему (5.4) получим формулы для определения неизвестных параметров регрессии a и b (или *МНК– оценок*):

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

(5.5)

или

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2} \quad ; \quad a = \bar{y} - b\bar{x}\ . \qquad \textbf{(5.6)}$$

Преобразуя формулу (5.5) для нахождения параметра b , получим:

$$b = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{x^2} - \left(\overline{x}\right)^2} \quad ; \qquad a = \overline{y} - b\overline{x}, \tag{5.6 (a)}$$

где $\quad \overline{x} = \frac{\sum x}{n}; \quad \overline{y} = \frac{\sum y}{n}; \quad \overline{xy} = \frac{\sum xy}{n}; \quad \overline{x^2} = \frac{\sum x^2}{n}$.

Уравнение регрессии *Y* по *X* можно также записать в виде:

$$y = b\left(x - \overline{x}\right) + \overline{y} \,. \tag{5.7}$$

Угловой коэффициент b в уравнении регрессии называется *выборочным коэффициентом регрессии Y по X* . Обозначим его b_{yx} ,

тогда уравнение регрессии можно записать в виде:

$$y = b_{yx}\left(x - \overline{x}\right) + \overline{y} \quad \text{или} \quad y - \overline{y} = b_{yx}\left(x - \overline{x}\right). \tag{5.8}$$

!

Коэффициент регрессии *Y* по *X* показывает на сколько единиц в *среднем* изменяется переменная *Y* при увеличении переменной *X* на одну единицу.

Из уравнений (5.8) видно, что прямая регрессии проходит через точку $\left(\overline{x}; \overline{y}\right)$, которая является центром тяжести системы данных точек.

?

Что понимается под центром тяжести системы?

Аналогично можно получить уравнение прямой регрессии *X* по *Y* :

$$x - \overline{x} = b_{xy}\left(y - \overline{y}\right) \,. \tag{5.9}$$

В формуле (5.9) b_{xy} – *выборочный коэффициент регрессии X по Y:*

$$b_{xy} = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{\overline{y^2} - \left(\overline{y}\right)^2} \,. \tag{5.10}$$

Формулы для нахождения коэффициентов регрессии можно переписать следующим образом:

$$b_{yx} = \frac{\mu}{S_x^{\ 2}}; \qquad b_{xy} = \frac{\mu}{S_y^{\ 2}}; \tag{5.11}$$

где:

$\mu = \overline{xy} - \overline{x} \cdot \overline{y}$ – *выборочная ковариация*,

$S_x^{\ 2} = \overline{x^2} - \left(\overline{x}\right)^2$ – *выборочная дисперсия* по переменной *X*,

$S_y^{\ 2} = \overline{y^2} - \left(\overline{y}\right)^2$ – *выборочная дисперсия* по переменной *Y*.

! В уравнениях регрессий b_{yx} и $\frac{1}{b_{xy}}$ – угловые коэффициенты линий регрессий, которые пересекаются в точке $(\bar{x}; \bar{y})$.

Предположим, что зависимость между *X* и *Y* задана корреляционной таблицей 5.2. Тогда в формулах для нахождения параметров уравнений регрессий появляются *веса*, задающие число одинаковых слагаемых:

$$\overline{x} = \frac{\sum_{i=1}^{k} x_i \cdot n_i}{n}; \quad \overline{y} = \frac{\sum_{j=1}^{m} y_j \cdot l_j}{n};$$

$$\overline{xy} = \frac{\sum_{i=1}^{k}\sum_{j=1}^{m} x_i \cdot y_j \cdot n_{ij}}{n}; \quad \overline{x^2} = \frac{\sum_{i=1}^{k} x_i^{\ 2} \cdot n_i}{n}. \tag{5.12}$$

Эмпирический коэффициент корреляции и его свойства

Предположим, что наблюдения проводят над системой двух случайных величин X и Y, связанных линейной зависимостью. Для оценки тесноты связи между ними по результатам выборки вычисляют статистику, называемую *коэффициентом корреляции*.

Коэффициент корреляции вычисляют по формулам:

$$r = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \sum_{i=1}^{n}(y_i - \bar{y})^2}}; \qquad \textbf{(5.13)}$$

$$r = \frac{\overline{xy} - \bar{x} \cdot \bar{y}}{S_x S_y} \quad \text{или} \quad r = \frac{\mu}{S_x S_y}. \qquad \textbf{(5.14)}$$

В формулах (5.14) S_x, S_y – выборочные средние квадратические отклонения соответственно по переменным X и Y, определяемые по формулам:

$$S_x = \sqrt{S_x^{\ 2}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n}} = \sqrt{\overline{x^2} - \left(\bar{x}\right)^2};$$

$$S_y = \sqrt{S_y^{\ 2}} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \bar{y})^2}{n}} = \sqrt{\overline{y^2} - \left(\bar{y}\right)^2}.$$

Если известны коэффициенты регрессии b_{yx} и b_{xy}, то коэффициент корреляции вычисляется по формуле:

$$r = \pm\sqrt{b_{yx} \cdot b_{xy}}\ . \qquad \textbf{(5.15)}$$

$\boxed{!}$ Коэффициент корреляции есть среднее геометрическое коэффициентов регрессии и имеет их знак.

Для практических расчётов иногда используют формулу, по которой r находят непосредственно из данных наблюдений, то есть на величине коэффициента корреляции не скажутся округления данных, связанных с расчётом средних и отклонений от них:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \left(\sum_{i=1}^{n} x_i\right)\cdot\left(\sum_{i=1}^{n} y_i\right)}{\sqrt{n\sum_{i=1}^{n} x_i^{\,2} - \left(\sum_{i=1}^{n} x_i\right)^2}\cdot\sqrt{n\sum_{i=1}^{n} y_i^{\,2} - \left(\sum_{i=1}^{n} y_i\right)^2}}. \qquad \textbf{(5.16)}$$

- Если выборочные данные заданы в виде корреляционной таблицы, то при вычислении коэффициента корреляции учитывают кратность значений (см. (5.12)).

Свойства коэффициента корреляции:

1.Значения коэффициента корреляции изменяются на отрезке [-1;1]:

$$-1 \le r \le 1 .$$

2. Чем больше r, тем теснее связь между изучаемыми признаками.

- В зависимости от того насколько $|r|$ близок к 1, различают связь *слабую, умеренную, заметную, достаточно тесную, тесную и весьма тесную.*

3. Если $|r| = 1$, то корреляционная связь становится функциональной.

4. Если $r = 0$, то между изучаемыми признаками нет линейной корреляционной зависимости, но не исключается существование какого либо другого вида корреляционной зависимости (параболической, гиперболической, показательной и т.д.).

5. Если $r > 0$, то связь между признаками *X, Y* прямая; если $r < 0$, то эта связь обратная.

- Выборочный коэффициент корреляции r является оценкой коэффициента корреляции ρ генеральной совокупности и поэтому также служит для измерения линейной связи между количественными признаками.

Если выборка имеет достаточно большой объём $(n \geq 50)$ и репрезентативна, то интервальная оценка коэффициента корреляции имеет вид:

$$r - t_{1-\alpha}\frac{\sqrt{1-r^2}}{\sqrt{n-2}} \leq \rho \leq r + t_{1-\alpha}\frac{\sqrt{1-r^2}}{\sqrt{n-2}}. \qquad \textbf{(5.17)}$$

- В двойном неравенстве (5.17) α – это уровень значимости, $\Phi(t_{1-\alpha}) = 1-\alpha = \gamma$ – доверительная вероятность. Значение $t_{1-\alpha}$ находят по таблице значений интегральной функции Лапласа (*таблица* 3 *Приложения* 3).

5.2. Примеры решения типовых задач

Задача 5.1. Изучается зависимость себестоимости единицы изделия (*Y*, тыс. руб.) от величины выпуска продукции (*X*, шт.) по группам предприятий за отчётный период. Экономист обследовал пять предприятий и получил следующие данные:

X	2	3	4	5	6
Y	1,9	1,7	1,8	1,6	1,4

Найти линейную регрессию *Y* на *X* и выборочный коэффициент корреляции.

Решение.

Построим данные точки (*x; y*) в системе координат (рис. 5.1).

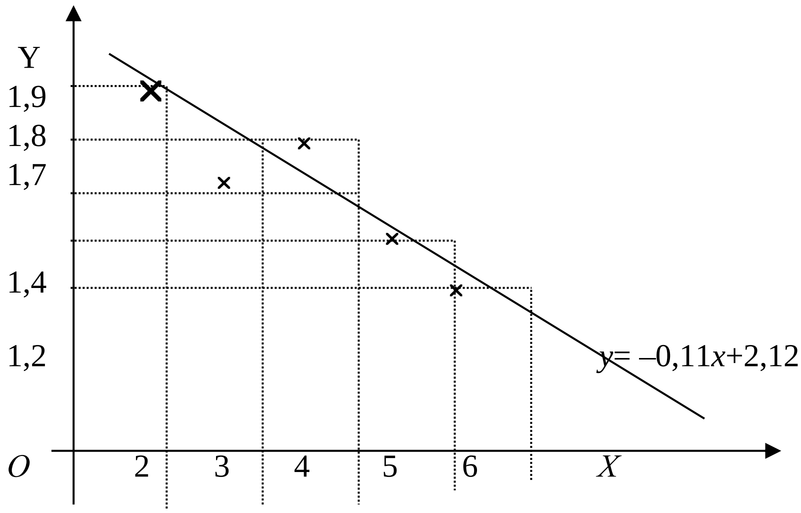


Рис. 5.1

Уравнение регрессии ищем в виде: $y = a + bx$, параметры b и a можно найти по одной из формул ((5.5), (5.6), (5.6*a*)). Так как данные при расчётах не требуют округления, то будем использовать формулу (5.6):

$$b = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}; \qquad a = \bar{y} - b\bar{x}.$$

Обозначим: $\Delta x_i = x_i - \bar{x}; \quad \Delta y_i = y_i - \bar{y}$, тогда имеем:

$$b = \frac{\sum_{i=1}^{n}\Delta x_i \Delta y_i}{\sum_{i=1}^{n}\Delta x_i^{\ 2}}.$$

Для вычислений удобно составить расчетную таблицу:

№	x_i	y_i	Δx_i	Δy_i	Δx_i^2	Δy_i^2	$\Delta x_i \Delta y_i$	$\hat{y}_i$	Δ_i	Δ_i^2
1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
1	2	1,9	-2	0,22	4	0,0484	-0,44	1,9	0	0
2	3	1,7	-1	0,02	1	0,0004	-0,02	1,79	-0,09	0,0081
3	4	1,8	0	0,12	0	0,0144	0	1,68	0,12	0,0144
4	5	1,6	1	-0,08	1	0,0064	-0,08	1,57	0,03	0,0009
5	6	1,4	2	-0,28	4	0,0784	-0,56	1,46	-0,06	0,0036
$\sum$	20	8,4	-	-	10	0,1480	-1,10	-	-	0,0270

Найдём среднюю выборочную по *X* и по *Y*. Для этого используем суммы, записанные в последней строке первого и второго столбцов:

$$\bar{x} = \frac{\sum x}{n} = \frac{20}{5} = 4; \quad \bar{y} = \frac{\sum y}{n} = \frac{8{,}4}{5} = 1{,}68.$$

Заполняем таблицу с четвёртого по восьмой столбец. В столбцах 6,7,8 подсчитываем суммы. Суммы, полученные по столбцам 6 и 8, подставляем в формулу для нахождения b:

$$b = \frac{-1{,}10}{10} = -0{,}11; \quad a = 1{,}68 - (-0{,}11) \cdot 4 = 2{,}12.$$

Тогда искомое уравнение регрессии: <u>*y=2,12 - 0,11x*</u>, где $b_{yx} = -0{,}11$ – коэффициент регрессии *Y* по *X*.

Теперь по уравнению регрессии рассчитаем теоретические значения $\hat{y}_i$ и заполним столбец 9. Например: $y_1 = 2{,}12 - 0{,}11 \cdot 2 = 1{,}9$.

Далее найдём отклонения наблюдаемых значений от теоретических:

$\Delta_i = y_i - \hat{y}_i$. Затем определим квадраты этих отклонений (то есть заполним столбцы 10 и 11 расчетной таблицы). Заметим, что в последнем столбце получили сумму квадратов отклонений, равную Δ_i^2=0,027. Так как в сравнении с данными число мало, следовательно выполнено условие МНК.

Коэффициент корреляции вычислим по формуле (5.13):

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

Итак, получаем: $r = \dfrac{-1{,}1}{\sqrt{10 \cdot 0{,}148}} = -0{,}904$.

Вывод: связь между признаками весьма тесная, обратная, т.е. чем больше продукции выпущено, тем меньше её себестоимость.

Ответ. *y=2,12 - 0,11x*, *r* = –0,904.

Задача 5.2. В таблице приведены данные о количестве внесённых удобрений *X* (ц/га) и урожайность *Y*(ц/га) на 100га пахотной земли:

X /*Y*	10	12	14	16	18	20	n_i
10	9	4	1				14
30	1	10	9	3			23
50		2	6	14	6		28
70			1	10	18	6	35
n_j	10	16	17	27	24	6	100

Требуется:

1.Найти коэффициент корреляции, сделать вывод о тесноте связи.

2.Составить уравнения прямых регрессии *Y* на *X* и *X* на *Y*, построить графики.

3. Определить величину среднего урожая, если внести по 80ц удобрений на гектар.

4. Найти, сколько надо внести удобрений, чтобы получить урожай 15ц с гектара.

Решение.

1.Для нахождения коэффициента корреляции используем формулу (5.14):

$$r = \frac{\overline{xy} - \overline{x} \cdot \overline{y}}{S_x \cdot S_y} \quad \text{или} \quad r = \frac{\mu}{S_x \cdot S_y}.$$

Найдём средние выборочные значения и средние квадратические отклонения переменных *X* и *Y* :

$$\overline{x} = \frac{1}{100}(10 \cdot 14 + 30 \cdot 23 + 50 \cdot 28 + 70 \cdot 35) = 46{,}80;$$

$$\bar{y} = \frac{1}{100}\left(10 \cdot 10 + 12 \cdot 16 + 14 \cdot 17 + 16 \cdot 27 + 18 \cdot 24 + 20 \cdot 6\right) = 15{,}14;$$

$$S_x{}^2 = \overline{x^2} - \left(\bar{x}\right)^2; \quad S_x = \sqrt{S_x{}^2}; \quad S_y{}^2 = \overline{y^2} - \left(\bar{y}\right)^2; \quad S_y = \sqrt{S_y{}^2}\ .$$

$$S_x{}^2 = \frac{1}{100}\left(10^2 \cdot 14 + 30^2 \cdot 23 + 50^2 \cdot 28 + 70^2 \cdot 35\right) - \left(46{,}80\right)^2 = 445{,}76;$$

Таким образом: $S_x = \sqrt{445{,}76} \approx 21{,}113$.

$$S_y{}^2 = \frac{1}{100}\left(10^2 \cdot 10 + 12^2 \cdot 16 + 14^2 \cdot 17 + 16^2 \cdot 27 + 18^2 \cdot 24 + 20^2 \cdot 6\right) -$$

$-\left(15{,}14\right)^2 = 8{,}0204.$ Значит: $S_y = \sqrt{8{,}0204} \approx 2{,}832.$

Чтобы найти ковариацию μ вычислим величину $\overline{xy}$.

$$\overline{xy} = \frac{1}{100}\left(10 \cdot 10 \cdot 9 + 10 \cdot 30 \cdot 1 + 10 \cdot 12 \cdot 4 + 30 \cdot 12 \cdot 10 + 50 \cdot 12 \cdot 2 + 14 \cdot 10 \cdot 1 + 30 \cdot 14 \cdot 9 + \right.$$
$$+ ? \cdot 50 \cdot 6 + 70 \cdot 14 \cdot 1 + 16 \cdot ? \cdot 3 + 50 \cdot 16 \cdot 14 \cdot + ? \cdot ? \cdot 10 + 18 \cdot 50 \cdot 6 +$$
$$\left. + 18 \cdot ? \cdot 18 + ? \cdot 70 \cdot ? \quad \right) = 759.$$

Тогда получаем, что $\mu = \overline{xy} - \bar{x} \cdot \bar{y} = 759 - 46{,}80 \cdot 15{,}14 = 50{,}448.$

Следовательно, коэффициент вариации $r = \dfrac{50{,}448}{21{,}113 \cdot 2{,}832} \approx 0{,}844.$

Вывод: связь между признаками тесная, прямая.

2. Найдём коэффициенты регрессии, используя формулы (5.11):

$$b_{yx} = \frac{\mu}{S_x{}^2} = \frac{50{,}448}{445{,}76} \approx 0{,}113; \qquad b_{xy} = \frac{\mu}{S_y{}^2} = \frac{50{,}448}{8{,}0204} \approx 6{,}290.$$

Составим теперь уравнения регрессий:

а) уравнение регрессии *Y* на *X*:

$$y-\bar{y}=b_{yx}(x-\bar{x}) \Rightarrow y-15{,}14=0{,}113(x-46{,}8) \Rightarrow$$
$$\Rightarrow\ y=0{,}113x+9{,}852.$$

б) уравнение регрессии X на Y:

$$x-\bar{x}=b_{xy}(y-\bar{y}) \Rightarrow x-46{,}80=6{,}290(y-15{,}14) \Rightarrow$$
$$\Rightarrow x=6{,}290y-48{,}431.$$

Графики найденных регрессий приведены на рис. 5.2.

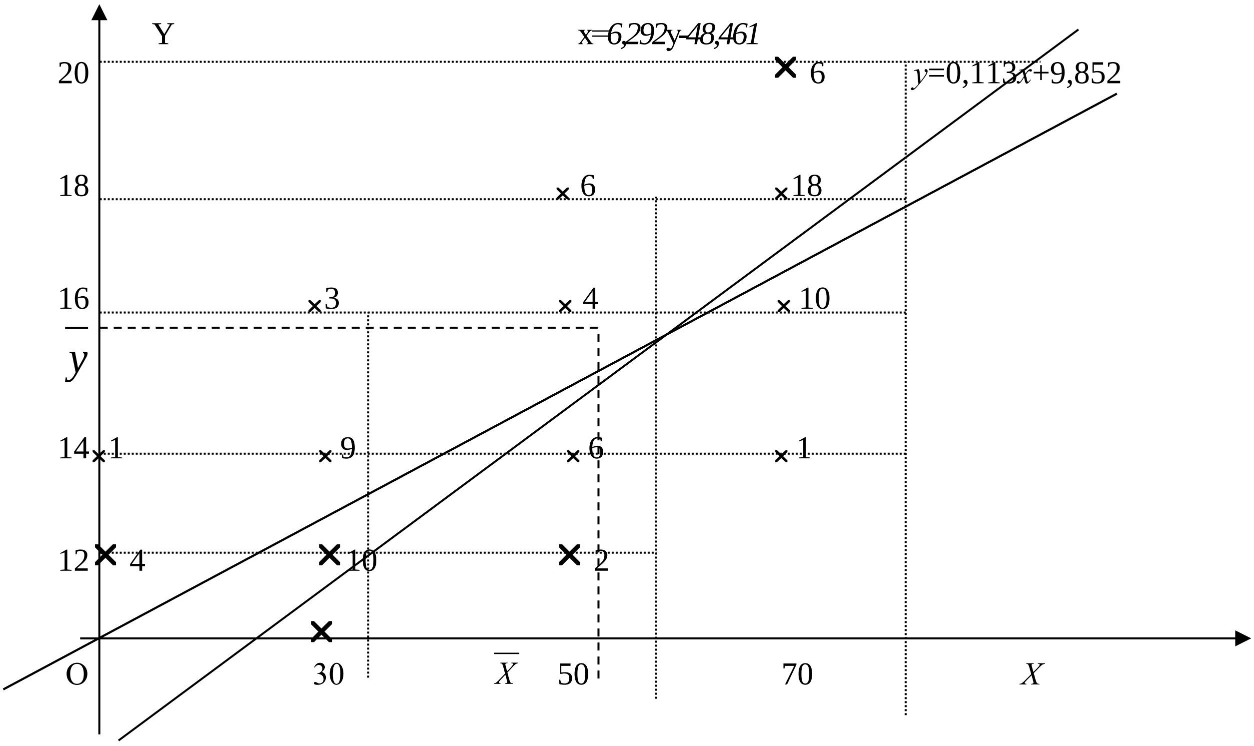


Рис.5.2

3.Чтобы найти урожайность, используя уравнение регрессии Y на X, найдём значение y при x=80: $y=0{,}113\cdot 80+9{,}852=18{,}892\approx 19$ ($ц/га$).

Значит, если будет внесено на 1 гектар 80 центнеров удобрений, то получим урожай примерно 19 центнеров с гектара.

4.Требуемое количество удобрений найдем, используя уравнение регрессии X на Y: $x=6{,}290\cdot 15-48{,}431=45{,}919\approx 46$. Для достижения урожайности в 15 ц/га необходимо внести 46 ц/га удобрений.

Ответ. 1. 0,844; **2.** $y=0{,}113x+9{,}852$; $x=6{,}290y-48{,}431$;
3. 19 ц/га; **4.** 46 ц/га.

- Для упрощения расчётов коэффициента корреляции и параметров уравнения регрессии используют упрощенный способ: метод условных вариантов (аналогично тому, как вычислялись характеристики вариационного ряда). Переходят к новым вариантам по следующим формулам:

$$u_i = \frac{x_i - c_1}{h_1}; \quad v_j = \frac{y_j - c_2}{h_2}; \tag{5.18}$$

где: c_1, c_2 – условные нули (выбираемые значения),

h_1, h_2 – разности между соседними значениями соответственно X и Y.

Для обратного перехода используют выражения:

$$x_i = h_1 u_i + c_1; \quad y_j = h_2 v_j + c_2;$$

$$\bar{x} = h_1 \cdot \bar{u} + c_1; \quad \bar{y} = h_2 \bar{v} + c_2;$$

$$S_x = h_1 S_u; \quad S_y = h_2 S_v.$$

где: $\bar{u}, \bar{v}$ – средние значения условных вариант,

S_u, S_v – средние квадратические отклонения условных вариант.

Выборочный коэффициент корреляции вычисляем по формуле:

$$r = \frac{\sum_{i=1}^{k} u_i \cdot U_i - n\bar{u}\bar{v}}{nS_u S_v} = \frac{\sum_{j=1}^{m} v_j \cdot V_j - n\bar{u}\bar{v}}{nS_u S_v},$$

где $U_i = \sum_{j=1}^{m} n_{ij} \cdot v_j; \quad V_j = \sum_{i=1}^{k} n_{ij} \cdot u_i.$

5.3. Задачи для самостоятельной работы

94.По исходным данным графически определить характер зависимости. Найти уравнение прямой регрессии Y на X и построить эту прямую.

X	1	1,5	3	4,5	5
Y	1,25	1,4	1,5	1,75	2,25

95.Опрос случайно выбранных 10 студентов вуза позволяет выявить зависимость между средним баллом по результатам предыдущей сессии и числом часов в неделю, затраченных студентом на самостоятельную подготовку. Соответствующие данные представлены в таблице:

Средний балл	4,6	4,3	3,8	3,8	4,2	4,3	3,8	4,0	3,1	3,9
Число часов	25	22	9	15	15	30	20	30	10	17

Постройте график исходных данных и определите по нему характер зависимости. Составьте выборочное уравнение регрессии. Рассчитайте выборочный коэффициент корреляции. Дайте интерпретацию полученных результатов.

96.Готовая продукция фирмы за 6 лет соответственно составила (в млн руб.):

Год	1995	1996	1997	1998	1999	2000
Продукция	18	21	26	22	25	28

Для выявления тенденции на ближайший год, найти по этим данным прямую МНК и вычислить прогнозируемую прибыль на следующий год.

97. В таблице приведены идеальные данные о росте и весе людей среднего возраста, сохранивших «спортивный» вес:

Рост (x_i)	178	166	172	168	176
Вес (y_i)	76	64	70	66	74

Вычислив коэффициент корреляции, убедится, что он равен 1.

98.При исследовании корреляционной зависимости по данным 20 предприятий между капиталовложениями X (млн руб.) и выпуском продукции Y (млн руб.), получены следующие уравнения регрессий: $y=1,2x+2$ и $x=0,7y+2$. Требуется найти: 1) среднее значение капиталовложений и выпуска продукции; 2) коэффициент корреляции.

99. При исследовании корреляционной зависимости между ценой на нефть X и индексом нефтяных компаний Y, получены следующие данные: $\bar{x} = 16,2(\textit{ден.ед.})$; $\bar{y} = 4000(\textit{усл.ед.})$;

$S_x^{\ 2} = 4;\ S_y^{\ 2} = 500;\ \mu = 40$.

1) Составить уравнения регрессий; 2) используя соответствующее уравнение регрессии, найти среднюю величину индекса при цене на нефть 16,5 (ден. ед.).

100. Распределение 50 гастрономических магазинов области по уровню издержек обращения X (%) и годовому объёму товарооборота Y(млн.руб.) представлено в таблице:

X/Y	0,5-2,0	2,0-3,5	3,5-5,0	5,0-6,5	6,5-8	n_x
4-6	-	-	-	3	2	5
6-8	-	4	8	8	1	21
8-10	2	5	5	2	-	14
10-12	3	1	5	-	-	9
12-14	1	-	-	-	-	1
n_y	6	10	18	13	3	50

1)Построить корреляционное поле; 2)составить уравнение регрессии Y по X и построить линию регрессии; 3) найти коэффициент корреляции.

101. Получено следующее распределение заводов по основным фондам X и готовой продукции Y (млн.руб.):

X/Y	20	30	40	50	60	Итого
15	7	5	-	-	-	12
25	20	23	-	-	-	43
35	-	30	47	2	-	79
45	-	10	11	20	6	47
55	-	-	9	7	3	19
Итого	27	68	67	29	9	200

Требуется: а) вычислить коэффициенты регрессии Y по X и X по Y; б) найти коэффициент корреляции и решить вопрос о тесноте связи между рассматриваемыми признаками;

в) составить уравнения прямых регрессии и построить их.

102. С целью анализа взаимного влияния зарплаты и текучести рабочей силы на пяти однотипных фирмах с одинаковым числом работников, проведены измерения уровня месячной зарплаты *X* и числа уволившихся за год рабочих *Y:*

X	100	150	200	250	300
Y	60	35	20	20	15

1) Построить график исходных данных и определите по нему характер зависимости. 2) Найти уравнение регрессии Y на X и выборочный коэффициент корреляции. 3) Построить найденную линию регрессии.

103. На основе выборочных данных о деловой активности некоторых коммерческих структур, оценивается теснота связи между прибылью Y(тыс. руб.) и затратами X (коп.) на 1 рубль произведённой продукции. Соответствующие данные приведены в таблице:

X	96	77	77	89	81
Y	220	1070	1000	600	790

Для данной выборки вычислены следующие величины:

$$\bar{x} = 84; \quad S_x = 7{,}43; \quad \bar{y} = 736; \; S_y = 306; \quad r_{xy} = -0{,}986 \; .$$

Найти МНК– прямую $y = a + bx$, для системы точек $(x;y)$. Построить данные точки и МНК – прямую.

104. Имеются выборочные данные о глубине вспашки полей под озимые культуры X(см) и их урожайности Y(ц/га):

X	10	15	20	25	30
Y	5	10	16	20	24

Вычислить эмпирический коэффициент корреляции и на его основе сделайте вывод о наличии и виде зависимости между признаками. Составить уравнение этой зависимости. Сделать прогноз урожайности пшеницы при глубине вспашки 22 см.

105. При исследовании корреляционной зависимости между объёмом продукции X (единиц) и её себестоимостью Y(тыс. руб.) получено следующее уравнение регрессии Y по X: $y = -0{,}0004x + 4{,}22$. Составьте уравнение регрессии X по Y, если коэффициент корреляции $r = -0{,}8$, а средний объём продукции $\bar{x} = 3000\,единиц$.

106. Зависимость между суточной выработкой продукции Y (тонн) и величиной основных фондов X(млн руб.) для совокупности 50 выбранных предприятий задана уравнениями регрессии:
$y_x = 0{,}676x - 4{,}79$; $x_y = 0{,}810y + 16{,}70$. Найти коэффициент корреляции и доверительный интервал для коэффициента корреляции генеральной совокупности с надёжностью $\gamma = 0{,}95$.

107. Получено распределение предприятий по объёму продукции X и по её себестоимости Y(руб.):

X/Y	2	2,5	3	3,5	4	Итого
1000	-	-	-	2	3	5
2000	-	-	3	6	2	11
3000	-	4	6	3	-	13
4000	1	6	4	1	-	12
5000	6	3	-	-	-	9
Итого	7	13	13	12	5	50

Требуется:

а) построить корреляционное поле; б) вычислить коэффициенты регрессии Y по X и X по Y; в) вычислить коэффициент корреляции и решить вопрос о тесноте связи между признаками; г) составить уравнения прямых регрессий и построить эти прямые.

108. Уровень расходов по зарплате впроцентах за год по одному магазину в зависимости от выполнения плана товарооборота (в %) дан в таблице, где X – процент выполнения плана товарооборота, Y – уровень расходов по зарплате (в %) за месяц:

X	101	98	104	105	98	85	97	103	103	101	106	102
Y	6,61	6,73	6,59	6,55	6,71	6,75	6,74	6,52	6,6	6,61	6,55	6,58

Определить тесноту связи и составить уравнение регрессии y_x.

109. Некоторая компания провела рекламную акцию в магазинах с демонстрацией нового товара. Через 10 недель компания решила проанализировать эффективность этого вида рекламы, сопоставив еженедельные объёмы продаж с расходами на рекламу (тыс. руб.):

Объём продаж, тыс.руб.	72	76	78	70	68	80	82	65	62	90
Расходы на рекламу, тыс, руб.	5	8	6	5	3	9	12	4	3	10

1) Постройте график исходных данных и определите по нему характер зависимости.
2) Рассчитайте выборочный коэффициент корреляции.
3) Составьте уравнение линии регрессии и постройте её.

110. Признаки X и Y связаны корреляционной зависимостью. Получены уравнения регрессий Y на X и X на Y:

$$y=0{,}382x+7{,}23; \quad x=1{,}587y+7{,}15.$$

Найти коэффициент корреляции и сделать вывод о тесноте связи, найдите средние значения признаков $\bar{x}, \bar{y}$.

111. Зависимость между товарными запасами Y(в днях оборота) и товарооборотом X (в тыс. руб.) задана таблицей:

X	30	50	70	90	125	175	230	250	300	400
Y	122	106	92	98	94	90	85	82	81	78

Постройте график исходных данных, определите тип зависимости, составьте уравнение регрессии y_x и постройте линию регрессии. Определите тесноту связи.

112. В результате выборочных наблюдений получены значения признаков X и Y:

Y/X	5	10	15	20	25	30	n_y
30	1	5	-	-	-	-	6
40	-	5	3	-	-	-	8
50	-	-	9	40	2	-	51
60	-	-	4	11	6	-	21
70	-	-	-	4	7	3	14
n_x	1	10	16	55	15	3	100

1) Оценить тесноту линейной связи между признаками по данным выборки.

2) Найти зависимость между признаками X и Y в виде уравнений линейной регрессии Y поX и X по Y.

3)Построить графически наблюдаемые значения признаков и линии регрессии.

113.Дана корреляционная таблица, в которой указаны: X – удои от коровы за одну лактацию, Y – содержание жира в молоке. Найдите: а)коэффициент корреляции; б) уравнение прямой линии регрессии Y на X.

Y/X	2000-2400	2400-2800	2800-3200	3200-3600	3600-4000	4000-4400	4400-4800	4800-5200	n_y
3,6-3,7	-	1	-	-	-	-	-	-	1
3,7-3,8	1	2	1	-	-	-	-	-	4
3,8-3,9	1	1	3	-	-	-	-	-	5
3,9-4,0	-	-	5	2	2	-	-	-	9
4,0-4,1	-	1	1	-	1	2	-	1	6
4,1-4,2	-	1	-	-	-	1	1	-	3
4,2-4,3	-	-	-	1	-	-	-	-	1
4,3-4,4	-	-	-	-	1	-	-	-	1
n_x	2	6	10	3	4	3	1	1	30

Тест 5

1. Существуют понятия статистической и корреляционной зависимости. Какая из этих зависимостей шире?

а) корреляционная; б) эти понятия совпадают;

в) статистическая; г) эти понятия не связаны между собой.

2.Известны средние значения признаков X и Y: $\bar{x} = 4;\ \bar{y} = 2$, а также коэффициент прямой регрессии Y на X: b_{yx}=1,5. Тогда уравнение регрессии будет иметь вид … .

3.Даны уравнения регрессий Y на X и X на Y:

$y= 0,9x+4;$ $x= 0,4y+1.$ Коэффициент корреляции равен:

а)0,36; б)1,3; в) 0,5; г) 0,6.

4. Коэффициент корреляции принимает значения:

а) $0<r<1$; б) $|r| \le 1$; в) $r>1$; г) r– любое число.

Данные о прибыли, полученной в течение месяца, за последние пять месяцев, оказались следующими:

Месяц	январь	февраль	март	апрель	май
Прибыль	42	40	60	71	87

С помощью МНК по этим точкам строится прямая регрессии. Эта прямая для прибыли в марте даёт значение (определите это значение без построения прямой регрессии): а)55; б) 71; в) 45; г) 60.

6. Если случайные величины X и Y связаны линейной зависимостью $y=a+bx$ (где $b>0$), a –любое, то коэффициент корреляции равен:

а) –1; б) 1; в) 0; г) a.

7. В моменты времени t_1, t_2,…t_n проводятся наблюдения, их результаты записываются в таблицу:

t	t_1	t_2	t_3	t_4	…	t_n
y	y_1	y_2	y_3	y_4	…	y_n

Для того, чтобы изучить функциональную тенденцию изменения наблюдаемой величины во времени, следует

а) построить вариационный ряд; б) построить график;

в) построить прямую методом наименьших квадратов;

г) вычислить $\bar{y};\ S^2$.

8. Наблюдения проводятся над системой двух случайных величин X и Y, найдены коэффициенты регрессий b_{yx} и b_{xy}. Тогда коэффициент корреляции r равен: а) $r = b_{yx} \cdot b_{xy}$; б) $r = \pm\sqrt{b_{yx} \cdot b_{xy}}$;

в) $r = \dfrac{b_{yx}}{b_{xy}}$; г) $r = \sqrt{b_{yx} \cdot b_{xy}}$.

9. Для некоторых признаков X и Y, связанных корреляционной зависимостью, уравнения регрессий имеют вид: $y=0,5x+4$; $x=0,8y-2$. Тогда средние значения этих признаков $\bar{x} = \ldots;\quad \bar{y} = \ldots$.

10. Зависимость между двумя переменными задана таблицей:

X	1	2	3	4	5
Y	10	20	30	40	50

Коэффициент корреляции r и коэффициент регрессии b_{yx} соответственно равны….

Вопросы для самопроверки по главе 5

1.Поятие статистической и корреляционных зависимостей.

3.В чём состоит различие между функциональной и корреляционной зависимостью?

4.Основные задачи регрессионного и корреляционного анализа.

5.Что такое корреляционное поле, как оно строится?

6.Опишите, как составляется корреляционная таблица.

7.Что называется эмпирической линией регрессии?

8.В каком виде ищут уравнение линии регрессии?

9. В чём состоит сущность метода наименьших квадратов для определения параметров линии регрессии?

10. Нахождение коэффициентов регрессии, что они показывают?

11.Определение коэффициент корреляции, его свойства.

Приложение 1

Таблица 1

Выборка A (n = 79; начало первого интервала – 0; длина интервала –1)

2	4	2	4	3	3	3	2	0	6
1	2	3	2	2	4	3	3	5	1
7	4	3	4	2	3	2	3	3	1
4	3	1	4	5	3	4	2	4	5
3	6	4	1	3	2	4	1	3	1
0	5	6	4	7	4	1	3	3	0
0	2	4	3	2	2	3	3	1	3
3	1	1	2	3	1	4	3	1	

Таблица 2

Выборка B (n =200; начало первого интервала –59; длина интервала –2)

65	71	67	73	68	68	72	68	67	70
78	74	79	65	72	65	71	70	69	69
76	71	63	77	75	70	74	65	71	68
73	73	76	72	73	69	67	75	64	74
74	75	75	76	64	71	66	78	68	69
80	74	76	68	74	70	65	71	70	69
69	72	73	69	71	72	73	70	66	66
73	80	72	69	69	71	70	69	70	71
76	69	69	67	67	74	68	74	60	69
74	72	69	71	69	74	67	71	69	63
78	69	71	66	66	75	69	66	75	77
66	72	68	81	72	71	71	71	70	69
75	76	72	68	70	72	73	71	73	65
72	68	71	67	69	71	71	74	63	72
69	78	74	73	74	71	71	66	77	70
68	69	67	78	72	75	68	67	74	74
63	61	68	73	76	69	76	69	75	67
70	72	64	69	74	74	74	74	73	68
70	68	74	66	71	68	70	74	73	74
78	76	73	73	67	66	72	73	62	69

Таблица 3

Выборка C (n = 69; начало первого интервала – 0; длина интервала –1)

0	3	5	2	3	1	4	3	1	2
4	1	5	3	0	1	3	3	3	5
2	5	3	4	2	3	2	5	1	2
0	2	2	3	2	2	3	3	5	1
5	0	2	2	3	4	6	5	4	1
1	2	0	3	2	2	2	3	5	1
1	2	2	2	2	2	1	0	4	

Таблица 4

Выборка D (n = 181; начало первого интервала – 102; длина интервала –4)

135	133	124	132	104	152	1343	130	129	120
132	126	125	133	132	133	120	118	129	122
120	131	118	122	120	132	136	144	115	124
117	127	115	131	126	132	140	129	147	117
127	130	134	128	115	135	138	120	131	123
118	126	130	137	132	132	134	109	132	123
120	124	132	111	134	125	126	127	137	129
135	123	143	115	127	132	115	122	119	121
116	123	122	117	127	108	124	126	125	122
118	121	139	130	124	114	134	129	120	125
133	111	120	113	135	121	118	127	134	131
136	123	140	128	128	133	135	124	131	147
125	129	137	123	127	133	129	135	120	124
126	127	120	129	115	114	125	127	135	137
119	116	125	128	137	115	123	134	131	112
129	120	131	120	127	122	127	108	125	126
124	126	118	115	125	131	132	132	117	128
127	109	138	128	116	125	118	120	118	111
132									

Таблица 5

Количество реализованного мяса свиней (в центнерах) по 80 сельхозпредприятиям Приморского края в 2016 г.

5	34	9	58	4	36	58	2	3	98
27	7	9	6	6	119	11	68	118	15
3	32	96	4	3	5	15	7	4	81
55	1	23	123	19	6	12	19	230	117
7	5	23	21	2	12	48	28	3	11
25	31	21	44	2	16	2	8	11	21
6	283	67	19	14	51	23	50	22	4
37	83	9	51	4	31	12	4	9	5

Таблица 6

Валовой сбор зерновых (без кукурузы) в весе после доработки (в центнерах) по 70 сельхозпредприятиям Приморского края в 2006 г.

609	1341	2833	7153	2926	1558	1720	2428	713	5624
1146	1456	5602	865	16899	1380	1950	1030	959	5721
3678	736	5513	4969	14015	2777	7909	6278	6490	780
1037	7557	898	6571	7773	14161	1164	2460	8024	26241
1305	5742	6574	1406	6119	934	2145	386	5216	1836
4456	3267	24820	2653	1942	4316	18415	675	1420	1456
3340	8221	20451	548	5057	2701	2067	5789	844	12655

Таблица 7

Сбор зерновых (без кукурузы) с 1 гектара (ц) по 50 сельхозпредприятиям Приморского края в 2016 г.

8,4	4,4	7,4	8,2	13,2	5,9	3,1	6,7	6,3	6,6
5,2	8,4	6,7	5,0	7,9	4,3	7,7	5,9	9,0	3,9
3,1	3,8	7,0	4,6	3,9	2,9	6,6	6,3	2,0	5,3
6,4	6,9	16,0	6,7	9,2	7,6	5,9	6,6	7,9	9,6
3,9	2,4	9,9	9,5	7,5	8,5	3,6	6,5	11,6	2,8

О т в е т ы

Глава 1. Выборка и ее представление

1.

x_i	4	7	8	12
$\tilde{p}_i$	0,25	0,10	0,15	0,50

2.

x_i	0	1	5	10
$\tilde{p}_i$	31/54	14/54	7/54	2/54

4.

$$F^*(x) = \begin{cases} 0, & при\ x \le 1; \\ 0{,}04, & при\ 1 < x \le 3; \\ 0{,}24, & при\ 3 < x \le 7; \\ 0{,}32, & при\ 7 < x \le 9; \\ 0{,}8, & при\ 9 < x \le 12; \\ 1, & при\ x > 12. \end{cases}$$

5.

$$F^*(x) = \begin{cases} 0, & при\ x \le 0; \\ 0{,}06, & при\ 0 < x \le 2; \\ 0{,}1, & при\ 2 < x \le 4; \\ 0{,}12, & при\ 4 < x \le 6; \\ 0{,}3, & при\ 6 < x \le 8; \\ 0{,}59, & при\ 8 < x \le 10; \\ 0{,}7, & при\ 10 < x \le 12; \\ 0{,}8, & при\ 12 < x \le 14; \\ 0{,}97, & при\ 14 < x \le 16; \\ 1, & при\ x > 16. \end{cases}$$

8.

$$F^*(x) = \begin{cases} 0, & при\ x \le 2; \\ 0{,}1, & при\ 2 < x \le 5; \\ 0{,}4, & при\ 5 < x \le 7; \\ 0{,}6, & при\ 7 < x \le 8; \\ 1, & при\ x > 8. \end{cases}$$

9.

x	60	65	70	75	100	120
m	3	3	7	5	8	4
p	3/30	3/30	7/30	5/30	8/30	4/30

13.

x	37	38	39	40	41	42	43	44
m	1	3	5	8	12	9	5	2

`14

$$F^*(x) = \begin{cases} 0, & при\ x \le -2; \\ 0{,}0375, & при\ -2 < x \le 0; \\ 0{,}25, & при\ 0 < x \le 5; \\ 0{,}6, & при\ 5 < x \le 8; \\ 0{,}875, & при\ 8 < x \le 14; \\ 1, & при\ x > 14. \end{cases}$$

15.

$$F^*(x) = \begin{cases} 0, & при\ x \le 11; \\ 0{,}16, & при\ 11 < x \le 14; \\ 0{,}4, & при\ 14 < x \le 17; \\ 0{,}7, & при\ 17 < x \le 20; \\ 0{,}77, & при\ 20 < x \le 23; \\ 0{,}85, & при\ 23 < x \le 26; \\ 0{,}91, & при\ 26 < x \le 29; \\ 0{,}96 & при\ 29 < x \le 32; \\ 1, & при\ x > 32. \end{cases}$$

16.

x_i	2	5	7	8	11	13
$\tilde{p}_i$	0,10	0,09	0,21	0,25	0,30	0,05

17.

$x_i < X \le x_{i+1}$	1-5	5-9	9-13	13-17	17-21
m_i / hm	$25 \cdot 10^{-3}$	$50 \cdot 10^{-3}$	$125 \cdot 10^{-3}$	$30 \cdot 10^{-3}$	$20 \cdot 10^{-3}$

Глава 2. Числовые характеристики вариационного ряда (выборки). Элементы теории оценок

21. 1. а) $\overline{x}_{в} = 5{,}76$;

22. 3269;

23. 6,075;

24. 101;

25. 1) $m_2 = 50$;

2) $D_в$=167,29;

26. 1) $\overline{x}_{в} = 12{,}94$;

2) $\sigma \approx 0{,}052$;

3) $D_в \approx 0{,}0027$;

29. 1. а) $\overline{x}_{в} = 4$;

30. $\overline{x}_{в} = 2621$;

31. 5,1;

32. 6;

33. 1) m_3=20; 3) $D_в$=0,0344

34. $\overline{x}_{в} = 18{,}67$; σ^2=2,23; $\sigma \approx 1{,}49$

36. Для смеси №1: $\overline{x}_1 \approx 10{,}96$; $\sigma_1 = 4{,}4$; для смеси №2: $\overline{x}_2 \approx 15{,}04$; $\sigma_1 = 5{,}04$. Вывод: смесь №2 более эффективна.

38. 1) $\overline{x}_{в} = 10$;

2)$D_в$=2,5; 3) S^2=10/3;

39. S^2=4,89;

45. Для рабочего №1: $\overline{x}_1 \approx 50{,}15\,мм$; $\sigma_1 = 0{,}058\,мм$;

Для рабочего №2: $\overline{x}_2 \approx 50{,}21\,мм$; $\sigma_2 = 0{,}076\,мм$;

Для рабочего №3: $\overline{x}_3 \approx 50{,}17\,мм$; $\sigma_1 = 0{,}084\,мм$.

Вывод: лучше всех работает первый рабочий, затем – третий, а второй рабочий работает хуже всех.

46. До наладки: $\overline{x}_1 \approx 20{,}31\,мм$; $\sigma_1 = 0{,}42\,мм$;

После наладки: $\overline{x}_1 \approx 20{,}07\,мм$; $\sigma_1 = 0{,}31\,мм$.

Вывод: наладка улучшила работу станка. Средний размер детали стал ближе к центру допуска, рассеяние уменьшилось.

Глава 10. Интервальные оценки числовых характеристик генеральной совокупности

51. $1370 \le X_{ген.} \le 1550$;

52. *а) (49,15; 50,85), б)(48,9; 51,1), в) (48,7; 51,3);*

53. $396 < X_{ген.} < 404$;

54. *а) (11,58; 12,42),*
б) (11,45; 12,55),
в) (11,40; 12,60);

55. *а)(18,63;21,57),*
б)(18,14;22,06),
в)(19,12;21,08);

56. *(75,23; 121,17);*

57. *0,9931;*

58. *144;*

59. 24;

60. *(0,104; 0,146);*

61. 33;

62. (3,5; 4,67);

63. $7,83 < M(X) < 10,27$;
$2,87 < \sigma(X) < 4,68$;

64. а)(0,47; 0,53),
б)(0,0255; 0,0745),
в)(0,48; 0,52);

65. $49,98 \le \overline{X}_в \le 54,68$;
$2,83 \le \sigma \le 7,85$;

66. $0,013 < p < 0,027$;

67. $0,344 < p < 0,416$;

68. $0,906 < p < 0,994$;

69. $P = 0,8098$;

70. $n_{max.} = 9600$; $n = 1474$;

71. $P = 0,9768$;

72. *1) от 27,68% до 32,32%; 2)n=13978;* *3)* $n \le 16641$.

Глава 11. Статистические гипотезы. Проверка статистических гипотез

75. $k_{кр} = -1,8$. Основная гипотеза отвергается. Значит, фирма в рекламе завышает срок безотказной работы изделия.

76. а) $H_0: a = 21;\ H_1: a < 21;\quad T_{набл} = -6{,}32 \in (-\infty;-1{,}645).$

Время обслуживания клиента (21 мин) в качестве норматива опытными данными не подтверждается.

б) $H_0: a = 16;\ H_1: a < 16;\quad T_{набл} = -1{,}054 \notin (-\infty;-1{,}645).$

Время обслуживания клиента (16 мин) в качестве норматива не противоречит опытным данным.

80. $H_0: a = 60;\ H_1: a < 60;\quad T_{набл} = -10 \in (-\infty;-1{,}689).$

Утверждение поставщика не согласуется с опытными данными.

83. $H_0: a = 35;\ H_1: a \neq 35;\ T_{набл} = 27 \in (-\infty;-1{,}993) \cup (1{,}993;\infty).$

Станок требует подналадки.

85. $H_0: a = 6;\ H_1: a > 6;\ T_{набл} = 20 \in (1{,}96;+\infty).$

Предположение о том, что будет продано 7 т хлеба, не противоречит опытным данным.

86. $T_{набл}$=–1,909; $k_{кр}$=–1,833; гипотеза $H_о$ отвергается

89. а) $H_0: a = 1000;\ H_1: a \neq 100;\ T_{набл} = -21 \in (-\infty;-1{,}96) \cup (1{,}96;\infty).$

Следовательно, гипотеза H_0 не отвергается.

б) $H_0: a = 1000;\ H_1: a \neq 100;\ T_{набл} = -1{,}656 \notin (-\infty;-2{,}032) \cup (2{,}032;\infty).$

Следовательно, гипотеза H_0 не отвергается.

Содержание

Жуплей Ирина Викторовна

Савельева Екатерина Владимировна

Жуплей, И.В. Основы математической статистики для экономического бакалавриата: учеб. пособие [Электронный ресурс] / И.В. Жуплей, Е.В. Савельева – Электрон. текст. дан. - Уссурийск, 2017. – 153 с. – Режим доступа: http://de.primacad.ru

ЭЛЕКТРОННОЕ ИЗДАНИЕ

ФГБОУ ВО Приморская ГСХА

Адрес: 692510, г. Уссурийск, пр-т Блюхера, 44