Методические указания и задания к РГР

по теме

Математическая статистика

Оглавление

ВВЕДЕНИЕ

В окружающей нас жизни приходится сталкиваться с различными явлениями и фактами, наступление которых приписывается случаю, а сами явления и факты называются случайными. Но такое представление связано с единичными явлениями и фактами или с небольшим количеством одинаковых случаев.

Изучение закономерностей однородных массовых случайных явлений составляет предмет теории вероятностей и основанной на ней математической статистики. При этом изучаемые явления рассматриваются в абстрактной форме независимо от их конкретной природы. Только такой метод, характерный для всех отраслей математических знаний, и позволяет обоснованно устанавливать общие закономерности и положения, которые могут затем применяться уже к достаточно широкому классу явлений. Однако использование законов теории вероятностей на практике возможно при условии тщательной проверки соблюдения основных положений теории вероятностей и при правильной статистической обработке материалов, относящихся к изучаемым массовым явлениям.

Математическая статистика – раздел математики, в котором изучаются методы сбора, систематизации и обработки результатов наблюдений массовых случайных явлений для выявления существующих закономерностей.

Учебное пособие подготовлено в соответствии с утвержденной программой курса «Математика» и требованиями действующего Государственного образовательного стандарта высшего образования.

В состав учебного пособия входят: основные положения курса математической статистики, варианты контрольных работ, указания по выполнению контрольной работы, решения типовых задач, решенный вариант контрольной работы, список рекомендуемой литературы.

1. ВЫБОРКИ И ИХ ХАРАКТЕРИСТИКИ

1.1. Предмет математической статистики

Математическая статистика – раздел математики, в котором изучаются методы сбора, систематизации и обработки результатов наблюдений массовых случайных явлений для выявления существующих закономерностей.

Математическая статистика тесно связана с теорией вероятностей. Обе эти математические дисциплины изучают массовые случайные явления. При этом теория вероятностей выводит из математической модели свойства реального процесса, а математическая статистика устанавливает свойства математической модели, исходя из данных наблюдений (говорят «из статистических данных»).

Предметом математической статистики является изучение случайных величин (или случайных событий, процессов) по результатам наблюдений. Полученные в результате наблюдения (опыта, эксперимента) данные сначала надо каким-либо образом обработать: упорядочить, представить в удобном для обозрения и анализа виде. Это первая задача. Затем, это уже вторая задача, оценить, хотя бы приблизительно, интересующие нас характеристики наблюдаемой случайной величины. Например, дать оценку неизвестной вероятности события, оценку неизвестной функции распределения, оценку математического ожидания, оценку дисперсии случайной величины, оценку параметров распределения, вид которого неизвестен, и т.д.

Следующей, назовем ее условно третьей, задачей является проверка статистических гипотез, т.е. решение вопроса согласования результатов оценивания с опытными данными. Например, выдвигается гипотеза, что: а) наблюдаемая случайная величина подчиняется нормальному закону; б) математическое ожидание наблюдаемой случайной величины равно нулю; в) случайное событие обладает данной вероятностью и т.д.

Одной из важнейших задач математической статистики является разработка методов, позволяющих по результатам обследования выборки (т.е. части исследуемой совокупности объектов) делать обоснованные выводы о распределении признака (случайной величины X) изучаемых объектов по всей совокупности.

Результаты исследования статистических данных методами математической статистики используются для принятия решения в задачах планирования, управления, прогнозирования и организации производства, при контроле качества продукции, при выборе оптимального времени настройки и замены действующей аппаратуры и т.д., то есть для научных и практических выводов.

Говорят, что *«математическая статистика – это теория принятия решений в условиях неопределенности».*

1.2. Генеральная и выборочная совокупности

Пусть требуется изучить данную совокупность объектов относительно некоторого признака. Например, рассматривая работу диспетчера, можно исследовать: его загруженность, тип клиентов, скорость обслуживания, моменты поступления заявок и т.д. Каждый такой признак (и их комбинации) образует случайную величину, наблюдения над которой мы и производим.

Совокупность всех подлежащих изучению объектов или возможных результатов всех мыслимых наблюдений, производимых в неизменных условиях над одним объектом, называется *генеральной совокупностью*.

Определение. Генеральная совокупность – это случайная величина $X(\omega)$, заданная на пространстве элементарных событий Ω с выделенным в ней классом S подмножеств событий, для которых указаны их вероятности.

Зачастую проводить сплошное обследование, когда изучаются все объекты, например – перепись населения, трудно или дорого, экономически нецелесообразно, а иногда невозможно. В этих случаях наилучшим способом обследования является выборочное наблюдение: выбирают из генеральной совокупности часть ее объектов («выборку») и подвергают их изучению.

Выборочной совокупностью (выборкой) называется совокупность объектов, отобранных случайным образом из генеральной совокупности.

Определение. Выборка – это последовательность $X_1, X_2, ..., X_n$ независимых одинаково распределенных случайных величин, распределение каждой из которых совпадает с распределением генеральной случайной величины.

Число объектов (наблюдений) в совокупности, генеральной или выборочной, называется ее *объемом*; обозначается соответственно через N или n. Конкретные значения выборки, полученные в результате наблюдений (испытаний), называют *реализацией* выборки и обозначают строчными буквами $x_1, x_2, ..., x_n$.

Метод статистического исследования, состоящий в том, что на основе изучения выборочной совокупности делается заключение о всей генеральной совокупности, называется *выборочным*.

Для получения хороших оценок характеристик генеральной совокупности необходимо, чтобы выборка была *репрезентативной* (или *представительной*), то есть достаточно полно представлять изучаемые признаки генеральной совокупности. Условием обеспечения репрезентативности выборки, является согласно закону больших чисел, соблюдение случайности отбора, то есть все объекты генеральной совокупности должны иметь равные вероятности попасть в выборку.

Различают выборки с возвращением (*повторные*) и без возвращения (*бесповторные*). В первом случае отобранный объект возвращается в генеральную совокупность перед извлечением следующего; во втором - не возвращается. Заметим, что если объем выборки значительно меньше объема

генеральной совокупности, различие между повторной бесповторной выборками очень мало, его можно не учитывать.

В зависимости от конкретных условий для обеспечения репрезентативности применяют различные способы отбора: *простой*, при котором из генеральной совокупности извлекают по одному объекту; *типический*, при котором генеральную совокупность делят на «типические» части и отбор осуществляется из каждой части; *механический*, при котором отбор производится через определенный интервал; *серийный*, при котором объекты из генеральной совокупности отбираются «сериями», которые должны исследоваться при помощи сплошного обследования. На практике обычно пользуются сочетанием вышеупомянутых способов отбора.

Пример 1. Десять абитуриентов проходят тестирование по математике. Каждый из них может набрать от 0 до 5 баллов включительно. пусть X_k - количество баллов, набранных $k-$м ($k=1, 2,..., 10$) абитуриентом.

Тогда значения 0, 1, 2, 3, 4, 5 – все возможные количества баллов, набранных одним абитуриентом, - образуют генеральную совокупность. Выборка $X_1, X_2,..., X_{10}$- результат тестирования 10 абитуриентов. Реализациями выборки могут быть следующие наборы чисел: {5, 3, 0, 1, 4, 2, 5, 4, 1, 5}или {4, 4, 5, 3, 3, 1, 5, 2, 2, 5}, то есть все возможные комбинации десяти чисел от 0 до 5.

1.3. Статистическое распределение выборки

Пусть изучается некоторая случайная величина X. С этой целью над случайной величиной производится ряд независимых опытов (наблюдений). В каждом из этих опытов величина X принимает то или иное значение.

Пусть она приняла m_1 раз значение x_1, m_2 раз – значение x_2, … , m_k раз - значение x_k. При этом $m_1 + m_2 + ... + m_k = n$ - объем выборки. Значения x_1, x_2, … , x_k называются *вариантами* случайной величины X, а изменение этих значений *варьированием*.

Расположение выборочных наблюдаемых значений случайной величины (признака) в порядке неубывания называется *ранжированием* статистических данных.

Полученная таким образом последовательность $x_{(1)}$, $x_{(2)}$, … , $x_{(n)}$ значений случайной величины X (где $x_{(1)} \le x_{(2)} \le \ldots \le x_{(n)}$ и $x_{(1)} = \min\limits_{1 \le i \le n} X_i$, … , $x_{(n)} = \max\limits_{1 \le i \le n} X_i$) называется *вариационным рядом*.

Числа m_i, показывающие сколько раз встречаются варианты x_i в ряде наблюдений, называются частотами, а отношение их к объему выборки –

частостями или *относительными частостями* (обозначают p_i^* или w_i), то есть

$$p_i^* = \frac{m_i}{n}, \quad \text{где } n = \sum_{i=1}^{k} n_i .$$

Перечень вариантов и соответствующих им частот или частостей называется *статистическим распределением ряда* или *статистическим рядом*.

Различают дискретные и непрерывные статистические ряды.

Дискретным статистическим рядом называется ранжированная совокупность вариант x_i с соответствующими им частотами. Записывается дискретный ряд в виде таблицы. Первая строка содержит варианты, а вторая их частоты или частости.

Пример 2. В результате тестирования (см. пример 1) группа абитуриентов набрала баллы: 5, 3, 0, 1, 4, 2, 5, 4, 1, 5. Записать полученную выборку в виде статистического ряда.

Решение.

Случайная величина X - число набранных баллов является дискретной случайной величиной.

Вначале составим ранжированный вариационный ряд $x_{(1)}, x_{(2)}, \ldots, x_{(10)}$, то есть расположим числа (баллы) в порядке неубывания их величин:

0, 1, 1, 2, 3, 4, 4, 5, 5, 5.

Подсчитав частоту и частость вариантов $x_1 = 0$, $x_2 = 1$, $x_3 = 2$, $x_4 = 3$, $x_5 = 4$, $x_6 = 5$ получим статистическое распределение выборки (так называемый дискретный статистический ряд):

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

$\left(\sum_{i=1}^{6} n_1 = 10\right)$

или

x_i	0	1	2	3	4	5
p_i^*	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{2}{10}$	$\frac{3}{10}$

$\left(\sum_{i=1}^{6} p_i^* = 1\right)$.

В случае, когда число значений признака (случайной величины X) велико или признак является непрерывным (то есть когда случайная величина X может принять любое значение в некотором интервале), составляют *интервальный* статистический ряд. В первую строку таблицы статистического распределения вписывают частичные промежутки $(x_0, x_1]$, $(x_1, x_2]$, … , $(x_{k-1}, x_k]$, которые берут обычно одинаковыми по длине. Для определения величины интервала h можно использовать формулу Стерджесса:

$$h = \frac{x_{\max} - x_{\min}}{1 + 3{,}32 \lg n},$$

где $x_{\max} - x_{\min} = R$ - размах признака, то есть разность между наибольшим и наименьшим значениями признака, $1 + 3{,}32 \lg n = m$ - число интервалов. За начало первого интервала рекомендуется брать величину $x_{\text{нач}} = x_{\min} - \frac{h}{2}$. Во второй строчке статистического ряда вписывают количество наблюдений $m_i \ (i = \overline{1,k})$, попавших в каждый интервал.

Пример 3. Измерили рост (с точностью до 1 см) 30 наудачу отобранных студентов. Результаты измерений таковы:
178, 160, 154, 183, 155, 153, 167, 186, 163, 155, 157, 175, 170, 160, 159,
173, 182, 167, 171, 169, 179, 165, 156, 179, 158, 171, 175, 173, 164, 172.

Построить интервальный статистический ряд.

Решение.

Для удобства проранжируем полученные данные:
153, 154, 155, 155, 156, 157, 158, 159, 160, 163, 164, 165, 166, 167, 167,
169, 170, 171, 171, 172, 173, 173, 175, 175, 178, 179, 179, 182, 183, 186.

Очевидно, что рост студентов – непрерывная случайная величина. Для полученной выборки: $x_{\min} = 153$, $x_{\max} = 186$. По формуле Стерджесса, при $n = 30$, находим длину частичного интервала:

$$h = \frac{186 - 153}{1 + 3{,}32 \lg 30} = \frac{33}{1 + 3{,}32 \lg 30} \approx \frac{33}{5{,}907} \approx 5{,}59.$$

Примем $h = 6$. Тогда $x_{нач} = 153 - \frac{6}{2} = 150$.

Число интервалов: $m = 1 + 3{,}32 \lg 30 = 5{,}907 \approx 6$.

Исходные данные разбиваем на 6 интервалов: $(150{,}156]$, $(156{,}162]$, $(162{,}168]$, $(168{,}174]$, $(174{,}180]$, $(180{,}186]$.

Подсчитав число студентов (m_i), попавших в каждый из полученных промежутков получим интервальный статистический ряд:

$x_i - x_{i+1}$	150-156	156-162	162-168	168-174	174-180	180-186
Частота m_i	4	3	6	7	5	3
Частость p_i^*	0,13	0,17	0,20	0,23	0,17	0,10

1.4. Эмпирическая функция распределения

Эмпирической (статистической) функцией распределения называется функция $F^*(x)$, определяющая для каждого значения x относительную частоту события $X < x$. Следовательно, по определению:

$$F^*(x) = p^*\{X < x\}.$$

Для нахождения эмпирической функции распределения удобно $F^*(x)$ записать в виде:

$$F^*(x) = \frac{m_x}{n},$$

где n – объем выборки, m_x – число выборочных значений величины X, меньших x.

Эмпирическую функцию распределения можно задать таблично или графически.

Пример 4. Построить функцию $F^*(x)$, используя условия и результаты примера 2.

Решение.

Объем выборки по условию примера $n = 10$. Наименьшая варианта равна 0, значит $m_x = 0$ при $x \le 0$ (наблюдений меньше 0 нет). Тогда $F^*(x) = \frac{0}{10} = 0$. Если $0 < x \le 1$, то неравенство $X < x$ выполняется для варианты $x_1 = 0$, которая встречается 1 раз ($m_x = 1$), поэтому $F^*(x) = \frac{1}{10} = 0{,}1$ и т.д. Окончательно получаем:

$$F^*(x) = \begin{cases} 0, & \text{при } x \le 0, \\ 0{,}1, & \text{при } 0 < x \le 1, \\ 0{,}3, & \text{при } 1 < x \le 2, \\ 0{,}4, & \text{при } 2 < x \le 3, \\ 0{,}5, & \text{при } 3 < x \le 4, \\ 0{,}6, & \text{при } 4 < x \le 5 \\ 1, & \text{при } 5 < x. \end{cases}$$

График эмпирической функции распределения приведен на рисунке 1.

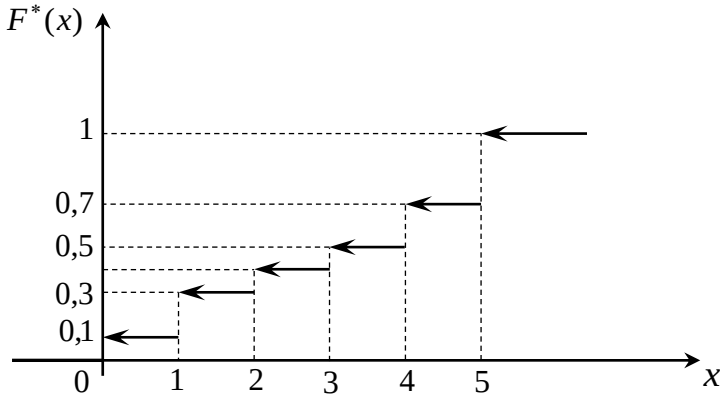

Рис. 1. Эмпирическая функция распределения дискретной случайной величины

В данном примере функция $F^*(x)$ есть выборочная функция распределения дискретной случайной величины и построена она по дискретному статистическому ряду.

Если случайная величина непрерывная и ее выборочные значения представлены в виде интервального статистического ряда, то выборочную функцию распределения строят иначе. Рассмотрим построение эмпирической функции распределения для интервального статистического ряда на примере.

Пример 5. Построить функцию $F^*(x)$, используя условия и результаты примера 3.

Решение.

Очевидно, что для $x \in (-\infty, 150]$ $F^*(x) = 0$, так как $m_x = 0$.

Используя результаты расчетов, представленных в таблице, подсчитаем на концах интервалов значения функции $F^*(x)$ в виде «наращенной относительной частоты»:

Рост	$[150,156)$	$[156,162)$	$[162,168)$	$[168,174)$	$[174,180)$	$[180,186)$
$F^*(x)$	0,13	0,30	0,50	0,73	0,90	1,00

Табличные значения не полностью определяют выборочную функцию распределения непрерывной случайной величины, поэтому при графическом изображении такой функции ее доопределяют, соединив точки графика, соответствующие концам интервала, отрезками прямой (рис.2):

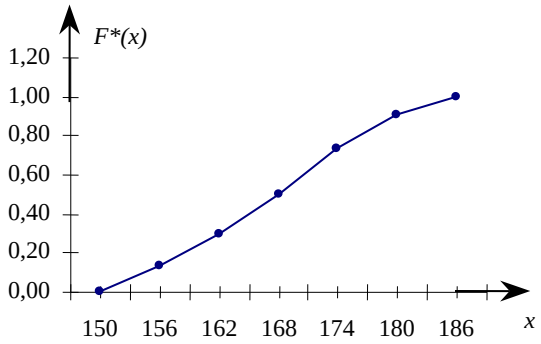

Рис. 2. Эмпирическая функция распределения
непрерывной случайной величины

1.5. Графическое изображение статистического распределения

Статистическое распределение изображается графически (для наглядности) в виде так называемых полигона и гистограммы. Полигон, как правило, служит для изображения дискретного статистического ряда (т.е. варианты отличаются на постоянную величину).

Полигоном частот называют ломаную, отрезки которой соединяют на плоскости точки с координатами (x_1, m_1), (x_2, m_2), … , (x_k, m_k); *полигоном частостей* – ломаную, соединяющую точки с координатами (x_1, p_1^*), (x_2, p_2^*), … , (x_k, p_k^*). Иногда полигон называют *многоугольником распределения*.

Варианты (x_i) откладываются на оси абсцисс, а частоты и соответственно частости - на оси ординат.

Пример 6. Пусть дана выборка в виде распределения частот:

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

$\left(\sum_{i=1}^{6} n_1 = 10\right)$

Построить полигон частостей.

Решение.

Статистический вариационный ряд можно записать в виде (см. пример 2):

x_i	0	1	2	3	4	5
p_i^*	0,1	0,2	0,1	0,1	0,2	0,3

$\left(\sum_{i=1}^{6} p_i^* = 1\right)$.

Полигон частостей для данного ряда имеет вид, изображенный на рис. 3:

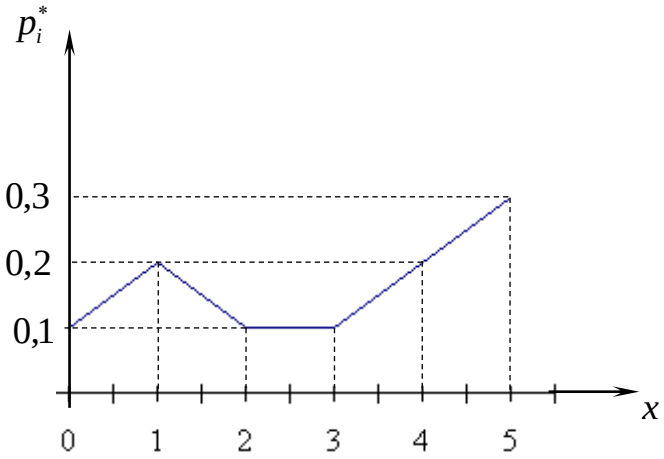

Рис.3. Полигон частостей

Полигон частостей является статистическим аналогом многоугольника распределения дискретной случайной величины.

Для непрерывно распределенного признака (то есть варианты могут отличаться одна от другой на сколь угодно малую величину) можно построить полигон частот, взяв середины интервалов в качестве значений признака $x_1, x_2, ..., x_k$. Однако чаще распределение непрерывного признака изображают графически в виде так называемой гистограммы.

Гистограммой частот (частостей) называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длины h, а высоты равны частотам или частостям соответствующих интервалов. Если соединить середины верхних оснований прямоугольников отрезками прямой, то можно получить полигон того же распределения.

Пример 7. Используя данные группировки промышленных предприятий по средней годовой стоимости основных производственных фондов:

Группы предприятий по стоимости ОПФ, млн.руб.	19,8-23,8	23,8-27,8	27,8-31,8	31,8-35,8	35,8-39,8
Число предприятий m_i	2	6	9	5	3

Требуется построить гистограмму частостей.

Решение.

Для построения гистограммы частостей, найдем p_i^*: так объем выборки $n = 25$, то $p_1^* = \frac{2}{25} = 0{,}08$; $p_2^* = \frac{6}{25} = 0{,}24$; $p_3^* = \frac{9}{25} = 0{,}36$; $p_4^* = \frac{5}{25} = 0{,}2$; $p_5^* = \frac{3}{25} = 0{,}12$. Гистограмма частостей изображена на рис.4:

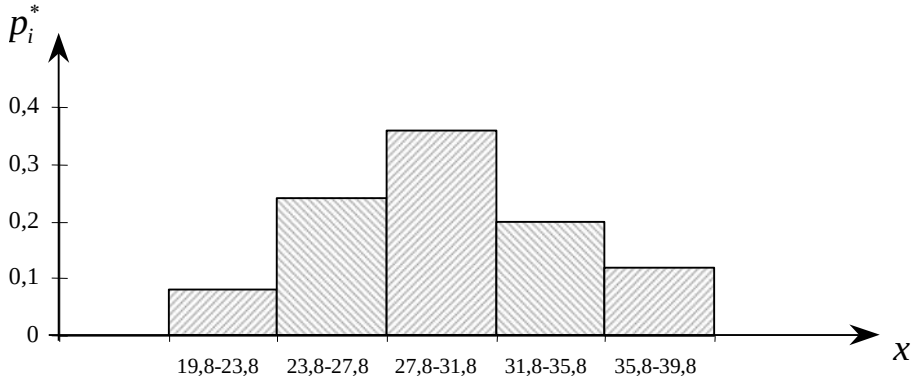

Рис. 4. Гистограмма частостей

Графическое изображение статистических распределений в виде полигона и гистограммы позволяет получить первоначальное представление о закономерностях, имеющих место в совокупности наблюдений.

2. СТАТИСТИЧЕСКОЕ ОЦЕНИВАНИЕ

2.1. Точечные оценки. Выборочная средняя и выборочная дисперсия

Оценки параметров генеральной совокупности, полученные на основании выборки, называются *статистическими*. Если статистическая оценка характеризуется одним числом, она называется *точечной*. К числу таких оценок относятся выборочная средняя и выборочная дисперсия.

Выборочная средняя определяется как среднее арифметическое полученных по выборке значений:

$$\overline{x}_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k} x_i \cdot n_i \ ,$$

где x_i – варианта выборки;

n_i – частота варианты;

n – объем выборки.

Выборочную среднюю можно записать и так:

$$\overline{x}_{\text{в}} = \sum_{i=1}^{k} x_i \cdot p_i^* \ ,$$

где $p_i^* = \frac{n_i}{n}$ – частость.

Выборочная средняя может обозначаться и без нижнего индекса: $\overline{x}$.

Отметим, что в случае интервального статистического ряда в качестве варианты x_i берут середины интервалов ряда, а в качестве n_i - частоты соответствующих интервалов.

Выборочной дисперсией называется среднее арифметическое квадратов отклонений значений выборки от выборочной средней $\bar{x}_в$:

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

или, что то же самое,

$$D_в = \sum_{i=1}^{k}\left(x_i - \bar{x}_в\right)^2 \cdot p_i^* .$$

Для расчетов может быть использована также формула:

$$D_в = \overline{x^2} - (\bar{x}_в)^2,$$

где $\overline{x^2}$ - выборочная средняя квадратов вариант выборки.

Выборочное *среднее квадратическое отклонение* выборки определяется формулой:

$$\sigma_в = \sqrt{D_в} .$$

Особенность выборочного среднего квадратического отклонения состоит в том, что оно измеряется в тех же единицах, что и изучаемый признак.

Статистическая оценка является случайной величиной и меняется в зависимости от выборки. Если математическое ожидание статистической оценки равно оцениваемому параметру генеральной совокупности, то такая оценка называется *несмещенной*, если не равно – то *смещенной*.

Выборочная средняя является оценкой математического ожидания случайной величины и представляет собой несмещенную оценку. Выборочная дисперсия оценивает дисперсию генеральной совокупности и является смещенной оценкой.

Для устранения смещенности выборочной дисперсии ее умножают на $n/(n-1)$ и получают величину:

$$S^2 = \frac{n}{n-1} D_в ,$$

которая называется несмещенной или *исправленной выборочной дисперсией*.

Величина

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

называется *исправленным выборочным средним квадратическим отклонением*.

Пример 8. Имеются данные о выручке в продовольственном магазине «Оазис» соответственно по месяцам (млн. руб.):

Месяц	1	2	3	4	5	6	7	8	9	10	11	12
Выручка	2,2	2,5	2,3	2,2	2,3	2,5	2,2	2,2	2,4	2,3	2,4	2,2

Найти выборочную среднюю и выборочную дисперсию.

Решение.

Построим сначала статистический ряд распределения:

Выручка, x_i	2,2	2,3	2,4	2,5
Частота, n_i	5	3	2	2

$\left(\sum_{i=1}^{4} n_1 = 12\right)$

Находим выборочную среднюю:

$$\bar{x}_{\text{в}} = \frac{1}{12}\sum_{i=1}^{4} x_i \cdot n_i = \frac{2{,}2 \cdot 5 + 2{,}3 \cdot 3 + 2{,}4 \cdot 2 + 2{,}5 \cdot 2}{12} = 2{,}31.$$

Для вычисления выборочной дисперсии используем формулу $D_{\text{в}} = \overline{x^2} - (\bar{x}_{\text{в}})^2$. Чтобы воспользоваться данной формулой найдем сначала $\overline{x^2}$:

$$\overline{x^2} = \frac{2{,}2^2 \cdot 5 + 2{,}3^2 \cdot 3 + 2{,}4^2 \cdot 2 + 2{,}5^2 \cdot 2}{12} = \frac{24{,}2 + 15{,}87 + 11{,}52 + 12{,}5}{12} = 5{,}34,$$

тогда $D_{\text{в}} = 5{,}34 - (2{,}31)^2 = 0{,}039.$

В качестве описательных характеристик вариационного ряда $x_{(1)}$, $x_{(2)}$, …, $x_{(n)}$ (или полученного из него статистического распределения выборки) используется медиана, мода, размах вариации (выборки).

Размахом вариации называется число:

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

где $x_{\max}$ - наибольшая, $x_{\min}$ - наименьшая варианты ряда.

Модой M_o вариационного ряда называется варианта, имеющая наибольшую частоту.

Медианой M_e вариационного ряда называется значение признака (варианта), приходящееся на середину ряда.

Если $n = 2k$ (то есть ряд $x_{(1)}$, $x_{(2)}$, …, $x_{(k)}$, $x_{(k+1)}$, $x_{(k+2)}$, …, $x_{(2k)}$ имеет четное число членов), то $M_e = \frac{x_{(k)} + x_{(k+1)}}{2}$. Если $n = 2k+1$ (то есть ряд имеет нечетное число членов), то $M_e = x_{(k+1)}$.

Пример 9. В результате тестирования (см. пример 2) группа абитуриентов набрала баллы: 5, 3, 0, 1, 4, 2, 5, 4, 1, 5. Найти характеристики выборки.

Решение.

Статистическое распределение выборки (так называемый дискретный статистический ряд) имеет вид:

x_i	0	1	2	3	4	5
n_i	1	2	1	1	2	3

$\left(\sum_{i=1}^{6} n_i = 10\right)$

Тогда:

$$\bar{x}_{\text{в}} = \frac{1}{10} \cdot (0 \cdot 1 + 1 \cdot 2 + 2 \cdot 1 + 3 \cdot 1 + 4 \cdot 2 + 5 \cdot 3) = 3,$$

$$D_{\text{в}} = \frac{1}{n} \sum_{i=1}^{k} (x_i - \bar{x}_{\text{в}})^2 \cdot n_i = \frac{1}{10} \cdot \left((0-3)^2 \cdot 1 + (1-3)^2 \cdot 2 + (2-3)^2 \cdot 1 + (3-3)^2 \cdot 1 + \right.$$

$$\left. + (4-3)^2 \cdot 2 + (5-3)^2 \cdot 3\right) = 3,2,$$

$$\sigma_{\text{в}} = \sqrt{D_{\text{в}}} = \sqrt{3,2} \approx 1,79, \qquad S^2 = \frac{n}{n-1} D_{\text{в}} = \frac{10}{9} \cdot 3,2 \approx 3,56,$$

$$S = \sqrt{S^2} = \sqrt{3,56} \approx 1,87, \qquad R = x_{\max} - x_{\min} = 5 - 0 = 5,$$

$M_o = 5$, так как 5 наиболее часто встречающаяся варианта,

$$M_e = \frac{x_{(5)} + x_{(6)}}{2} = \frac{3+4}{2} = 3,5.$$

Для непрерывно распределенного признака формулы для вычисления моды и медианы имеют вид:

$$M_o = x_{Mo} + h \cdot \frac{f_{Mo} - f_{(Mo-1)}}{\left(f_{Mo} - f_{(Mo-1)}\right) + \left(f_{Mo} + f_{(Mo-1)}\right)},$$

где x_{Mo} – начало модального интервального интервала, то есть интервала, имеющего наибольшую частоту,

f_{Mo} – частота модального интервального,

$f_{(Mo-1)}$ – частота интервала, предшествующего модальному,

$f_{(Mo+1)}$ – частота интервала, следующего за модальным,

h – интервал группировки;

$$M_e = x_{Me} + h \cdot \frac{\frac{n+1}{2} - S_{(Me-1)}}{f_{Me}},$$

где x_{Me} – начало медианного интервала, то есть интервала содержащего серединные значения вариационного ряда,

$S_{(Me-1)}$ – накопленная частота интервала, предшествующего модальному.

2.2. Метод моментов

При заданном виде закона распределения случайной величины X неизвестные параметры этого распределения можно оценить, то есть выразить как функцию вариант выборки.

Одним из методов нахождения точечных оценок неизвестных параметров заданного распределения является так называемый *метод моментов*.

Этот метод состоит в том, что приравниваются соответствующие теоретические и эмпирические моменты и из полученных уравнений находятся оценки параметров. В случае одного параметра в теоретическом распределении для его оценки достаточно составить одно уравнение. Если имеются два параметра в теоретическом распределении, то нужно приравнять соответственно два теоретических и эмпирических момента и т.д.

Для оценки двух параметров закона распределения запишем следующие равенства:

$$\nu_1 = M_1, \qquad \mu_2 = m_2,$$

где ν_1 - начальный момент первого порядка закона распределения случайной величины;

M_1 - эмпирический момент первого порядка;

μ_2 - центральный момент второго порядка закона распределения случайной величины;

m_2 - центральный эмпирический момент второго порядка.

Так как $\nu_1 = M_x$ - математическое ожидание случайной величины X, $\mu_2 = D_x$ - дисперсия случайной величины X, $M_1 = \bar{x}_в$, $m_2 = D_в$, то получаем два уравнения:

$$M_x = \bar{x}_в, \quad D_x = D_в.$$

Пример 10. Имеются данные за шесть месяцев об остатках вкладов населения на счетах некоторого коммерческого банка (млн. руб.):

Месяц	1	2	3	4	5	6
Остатки вкладов	20	24	25	28	27	32

Остатки вклада на первое число каждого месяца являются случайной величиной, для характеристики которой принят показательный закон распределения

$$f(x) = \lambda \cdot e^{-\lambda x} \qquad (x \geq 0).$$

Найти оценку параметра λ.

Решение.

Так как закон распределения содержит лишь один параметр λ, то для его оценки требуется составить одно уравнение.

Находим выборочную среднюю:

$$\overline{x}_{\text{в}} = \frac{20 + 24 + 25 + 28 + 27 + 32}{6} = 26.$$

Определяем математическое ожидание:

$$M_x = \int_0^{\infty} x \cdot f(x)\, dx = \lambda \int_0^{\infty} x \cdot e^{-\lambda x} dx.$$

Интегрируя по частям, получаем:

$$\int_0^{\infty} x \cdot e^{-\lambda x} dx = \left| \begin{array}{l} U = x; \quad dV = e^{-\lambda x} dx; \\ dU = dx; \quad V = -\dfrac{1}{\lambda} e^{-\lambda x} \end{array} \right| = -\frac{1}{\lambda} x e^{-\lambda x} \Big|_0^{\infty} + \frac{1}{\lambda} \int_0^{\infty} e^{-\lambda x} dx =$$

$$= -\frac{1}{\lambda} x e^{-\lambda x} \Big|_0^{\infty} - \frac{1}{\lambda^2} e^{-\lambda x} \Big|_0^{\infty} = -\frac{1}{\lambda} \left(x e^{-\lambda x} + \frac{1}{\lambda} e^{-\lambda x} \right) \Big|_0^{\infty} = -\frac{1}{\lambda} \left(-\frac{1}{\lambda} \right) = \frac{1}{\lambda^2}.$$

Тогда

$$M_x = \lambda \frac{1}{\lambda^2} = \frac{1}{\lambda},$$

Откуда

$$\frac{1}{\lambda} = \overline{x}_{\text{в}}.$$

Последнее равенство является приближенным, так правая часть его является случайной величиной. Таким образом, из полученного уравнения получается не точное значение λ, а его оценка λ^*:

$$\lambda^* = \frac{1}{\overline{x}_{\text{в}}} = \frac{1}{26}.$$

2.3. Метод максимального правдоподобия

Пусть $x_1,\ x_2,\ldots,\ x_n$ - выборка, полученная в результате проведения n независимых наблюдений за случайной величиной X. И пусть вид закона распределения величины X, например вид плотности $f(x,\theta)$, известен, но неизвестен параметр θ, которым определяется этот закон. Требуется по выборке оценить параметр θ.

Метод максимального правдоподобия, предложенный Р.Фишером, опирается на использование условий экстремума функции одной или нескольких случайных величин. В качестве такой функции принимают *функцию правдоподобия*.

Для дискретной случайной величины функция правдоподобия принимает вид

$$L = p(x_1,\theta)\cdot p(x_2,\theta)\cdot\ldots\cdot p(x_n,\theta),$$

где $x_1,\ x_2,\ldots,\ x_n$ – варианты выборки;

θ – параметр, для которого находится оценка;

$p(x_i,\theta) = P(X = x_i,\theta)$ – вероятность события $X = x_i$, зависящая от параметра θ.

Для непрерывных случайных величин функция правдоподобия имеет вид:

$$L = f(x_1,\theta)\cdot f(x_2,\theta)\cdot\ldots\cdot f(x_n,\theta),$$

где $f(x_i,\theta)$ – плотность распределения случай ной величины X.

Из определения следует, что чем больше значение функции $L(x,\theta)$, тем более вероятно (правдоподобнее) появление при фиксированном θ в результате наблюдений чисел $x_1,\ x_2,\ldots,\ x_n$.

За точечную оценку параметра θ, согласно методу максимального правдоподобия , берут такое его значение θ^*, при котором функция правдоподобия достигает максимума.

Так как функции $L(x,\theta)$ и $\ln L(x,\theta)$ достигают максимума при одном и том же значении θ, то вместо отыскания максимума функции $L(x,\theta)$ ищут (что проще) максимум функции $\ln L(x,\theta)$.

Таким образом, для нахождения оценки максимального правдоподобия надо:

1. решить уравнение правдоподобия
$$\left.\frac{d(\ln L(x,\theta))}{d\theta}\right|_{\theta=\theta^*} = 0;$$
2. отобрать то решение, которое обращает функцию $\ln L(x,\theta)$ в максимум. Удобно использовать вторую производную: если

$$\left.\frac{d^2\left(\ln L(x,\theta)\right)}{d\theta^2}\right|_{\theta=\theta^*} < 0,$$

то $\theta = \theta^*$ – точка максимума (достаточное условие).

Если оценке подлежат несколько параметров θ_1, θ_2,…, θ_n распределения, то оценки $\theta_1^*,...,\theta_2^*$ определяются решением системы уравнений правдоподобия:

$$\begin{cases} \dfrac{\partial(\ln L)}{\partial\theta_1} = 0, \\ ..., \\ \dfrac{\partial(\ln L)}{\partial\theta_n} = 0. \end{cases}$$

Чаще всего метод максимального правдоподобия используется при биномиальном, пуассоновском и показательном распределениях случайной величины.

В случае б и н о м и а л ь н о г о распределения

$$P_r(m) = C_r^m p^m (1-p)^{r-m},$$

где $P_r(m)$ – вероятность появления ровно m раз события A (случайной величины) в r испытаниях;

p – вероятность появления события A в одном испытании.

Величина p может рассматриваться как параметр.

Если проводится n опытов по r испытаний в каждом и фиксируется число появлений события (величины) в каждом испытании x_i, то при подстановке этого значения в формулу биномиального распределения получаем:

$$P_r(x_i, p) = C_r^{x_i} p^{x_i} (1-p)^{r-x_i}.$$

Тогда функция правдоподобия примет вид:

$$L = p_r(x_1, p)\cdot p_r(x_2, p)\cdot ...\cdot p_r(x_n, p).$$

После логарифмирования и приравнивания к нулю производной от $\ln L$ получаем выражение для оценки

$$p^* = \sum_{i=1}^{n} \frac{x_i}{nr}.$$

Если значения x_i встречаются n_i раз, то оценка параметра p примет вид:

$$p^* = \sum_{i=1}^{k} \frac{x_i n_i}{nr},$$

где $n = n_1 + n_2 + ... + n_k$ – число опытов по r испытаний в каждом.

В случае п у а с с о н о в с к о г о распределения

$$P_r(m) = \frac{\lambda^m}{m!} e^{-\lambda}$$

при подстановке вариант выборки получаем:

$$P_r(x_i, \lambda) = \frac{\lambda^{x_i}}{x_i!} e^{-\lambda}.$$

Составив функцию правдоподобия L, дифференцируя $\ln L$ и приравнивая его производную к нулю, находим оценку параметра λ в виде:

$$\lambda^* = \sum_{i=1}^{n} \frac{x_i}{n} = \bar{x}_{\text{в}}, \text{ или}$$

$$\lambda^* = \sum_{i=1}^{л} \frac{n_i x_i}{n} = \bar{x}_{\text{в}}.$$

В случае п о к а з а т е л ь н о г о распределения

$$f(x) = \lambda e^{-\lambda x} \qquad (x \geq 0)$$

функция правдоподобия для выборочных значений x_1, x_2,…, x_n примет вид:

$$L = \lambda e^{-\lambda x_1} \cdot \lambda e^{-\lambda x_2} \cdot \ldots \cdot \lambda e^{-\lambda x_n} = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}.$$

После преобразований получаем выражение для оценки параметра λ:

$$\lambda^* = \frac{n}{\sum_{i=1}^{n} x_i} = \frac{1}{\bar{x}_{\text{в}}}.$$

Пример 11. Партия стеклянных сувениров отправлена для реализации из Москвы в Иркутск в 1000 контейнерах. После поступления товара было выявлено количество разбитых изделий x_i в каждом контейнере n_i. Результаты представлены в таблице:

x_i	0	1	2	3	4
n_i	785	163	32	16	1

Считая, что число разбитых сувениров описывается законом Пуассона, найти точечную оценку параметра λ.

Решение.

Для нахождения точечной оценки параметра λ распределения Пуассона воспользуемся известной формулой

$$\lambda^* = \bar{x}_{\text{в}}.$$

Вычислим выборочную среднюю данного распределения:

$$\bar{x}_{\text{в}} = \sum_{i=1}^{л} \frac{n_i x_i}{n} = \frac{785 \cdot 0 + 163 \cdot 1 + 32 \cdot 2 + 16 \cdot 3 + 4 \cdot 4}{1000} = 0{,}29.$$

Тогда $\lambda^* = 0{,}29$.

2.4. Интервальное оценивание параметров

Точечные оценки неизвестного параметра θ хороши в качестве первоначальных результатов обработки наблюдений. Их недостаток в том, что неизвестно, с какой точностью они дают оцениваемый параметр. Для выборок небольшого объема вопрос о точности оценок очень существенен, так как между θ и θ^* может быть большое расхождение в этом случае. Кроме того, при решении практических задач часто требуется определить и надежность этих оценок. Тогда и возникает задача о приближении параметра θ не одним числом, а целым интервалом (θ_1^*, θ_2^*).

Оценка неизвестного параметра называется *интервальной*, если она определяется двумя числами – концами интервала.

Задачу интервального оценивания можно сформулировать следующим образом: по данным выборки построить числовой интервал (θ_1^*, θ_2^*), относительно которого с заранее выбранной вероятностью γ можно сказать, что внутри этого интервала находится точное значение оцениваемого параметра.

Интервал (θ_1^*, θ_2^*), в который с заданной вероятностью γ попадает истинное значение параметра θ, называется *доверительным интервалом*, а вероятность γ – *надежностью оценки* или *доверительной вероятностью*.

Величина γ выбирается заранее, ее выбор зависит от конкретно поставленной задачи. Так, степень доверия авиапассажира к надежности самолета, должна быть выше степени доверия покупателя к надежности телевизора. Надежность γ принято выбирать равной 0,9; 0,95; 0,99 или 0,999.

Доверительный интервал для *оценки математического ожидания* M_x случайной величины X с заданной надежностью γ в случае нормального распределения *при известной дисперсии* определяется на основе неравенств:

$$\bar{x}_в - t \cdot \frac{\sigma_x}{\sqrt{n}} < a = M_x < \bar{x}_в + t \cdot \frac{\sigma_x}{\sqrt{n}},$$

то есть это интервал

$$\left(\bar{x}_в - t \cdot \frac{\sigma_x}{\sqrt{n}},\ \bar{x}_в + t \cdot \frac{\sigma_x}{\sqrt{n}} \right),$$

где t – значение аргумента функции Лапласа, получаемое из таблиц, с учетом того, $\Phi(t) = \frac{\gamma}{2}$;

σ_x – известное среднее квадратическое отклонение или его оценка;

n – объем выборки.

Пример 12. Для проведения статистического исследования отобрано 25 промышленных предприятий отрасли. Средняя стоимость основных фондов предприятий оказалась равной 37 млн. руб. Предположим распределение стоимости основных фондов предприятий отрасли нормальным со средним квадратическим отклонением $\sigma_x = 4{,}1$, найти доверительный интервал для оценки математического ожидания с надежностью 0,95.

Решение.

Учитывая, что $\gamma = 0{,}95$ и $\Phi(t) = \frac{\gamma}{2}$, находим значение функции $\Phi(t) = 0{,}475$. По таблице значений функции Лапласа находим: $t = 1{,}96$. Тогда $t \cdot \frac{\sigma_x}{\sqrt{n}} = 1{,}96 \cdot \frac{4{,}1}{\sqrt{25}} = 1{,}6$. Следовательно, согласно формуле $\left(\bar{x}_в - t \cdot \frac{\sigma_x}{\sqrt{n}};\ \bar{x}_в + t \cdot \frac{\sigma_x}{\sqrt{n}} \right)$ доверительный интервал для математического ожидания будет:

$$(\bar{x}_в - 1{,}6;\ \bar{x}_в + 1{,}6);$$

$$(37 - 1{,}6;\ 37 + 1{,}6);$$

$$(35{,}4;\ 38{,}6).$$

Таким образом, с вероятностью 0,95 можно утверждать, что средняя стоимость основных фондов промышленных предприятий принадлежит интервалу: $(35{,}4;\ 38{,}6)$.

Для получения *оценки математического ожидания* нормально распределенной случайной величины *при неизвестной дисперсии* следует воспользоваться неравенством:

$$\bar{x}_в - t_\gamma \cdot \frac{S}{\sqrt{n}} < a = M_x < \bar{x}_в + t_\gamma \cdot \frac{S}{\sqrt{n}},$$

то есть это интервал

$$\left(\bar{x}_в - t_\gamma \cdot \frac{S}{\sqrt{n}},\ \bar{x}_в + t_\gamma \cdot \frac{S}{\sqrt{n}} \right),$$

где $S -$ исправленное среднее квадратическое отклонение случайной величины X, вычисленное по выборке: $S = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x}_в)^2}$;

$t_\gamma -$ находится по таблице квантилей распределения Стьюдента в зависимости от доверительной вероятности γ и числа степеней свободы $n - 1$.

Пусть X – нормально распределенная случайная величина с параметрами (a, σ). Пусть $\sigma -$ неизвестно, $\gamma -$ задана. Можно показать, что если

математическое ожидание a известно, то *доверительный интервал* для *среднего квадратического отклонения* σ имеет вид:

$$\left(\frac{\sqrt{n} \cdot S_0}{\chi_2}; \frac{\sqrt{n} \cdot S_0}{\chi_1} \right),$$

где $n-$ объем выборки, $S_0^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - a)^2$, а

$$\chi_1^2 = \chi^2_{\frac{1+\gamma}{2},n}; \ \chi_2^2 = \chi^2_{\frac{1-\gamma}{2},n}$$

являются квартилями χ^2 распределения с n степенями свободы, определяемые по таблице квартилей $\chi^2_{\alpha,n}$.

Если математическое ожидание a неизвестно, то доверительный интервал для неизвестного σ имеет вид:

$$\left(\frac{\sqrt{n-1} \cdot S}{\chi_2}; \frac{\sqrt{n-1} \cdot S}{\chi_1} \right),$$

где $S = \sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^{n} (x_i - \bar{x}_в)^2}$ – исправленное среднее квадратическое отклонение, квантили

$$\chi_1^2 = \chi^2_{\frac{1+\gamma}{2},n-1}; \ \chi_2^2 = \chi^2_{\frac{1-\gamma}{2},n-1}$$

определяются по таблице $\chi^2_{\alpha,k}$ при $k = n-1$ и $\alpha = \frac{1 \pm \gamma}{2}$ соответственно.

Пример 13. Для оценки параметра нормально распределенной случайной величины (средняя месячная зарплата работников предприятия, тыс. руб.)была сделана выборка объемом 30 единиц и вычислено исправленное среднее квадратическое отклонение $S = 1{,}5$. Найти доверительный интервал, покрывающий σ с вероятностью $\gamma = 0{,}90$.

Решение.

Имеем $n = 30$, $\gamma = 0{,}9$. По таблице $\chi^2_{\alpha,k}$ находим:

$$\chi_1^2 = \chi^2_{\frac{1+0{,}9}{2},30-1} = \chi^2(0{,}95; 29) = 17{,}7,$$

$$\chi_2^2 = \chi^2_{\frac{1-0{,}9}{2},30-1} = \chi^2(0{,}05; 29) = 42{,}6.$$

Доверительный интервал имеет вид:

$$\left(\frac{\sqrt{30-1} \cdot 1{,}5}{\sqrt{42{,}6}}; \frac{\sqrt{30-1} \cdot 1{,}5}{\sqrt{17{,}7}} \right)$$

или $1{,}238 < \sigma < 1{,}920$.

Таким образом с вероятностью $0{,}9$ можно утверждать, что среднее квадратическое отклонение заработной платы работников данного предприятия принадлежит интервалу $(1{,}238\,;1{,}920)$.

3. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

3.1. Задачи статистической проверки гипотез

Одна из часто встречающихся на практике задач, связанных с применением статистических методов, состоит в решении вопроса о том, должно ли на основании данной выборки быть принято или, наоборот отвергнуто некоторое предположение (гипотеза) относительно генеральной совокупности (случайной величины).

Например, новое лекарство испытано на определенном числе людей. Можно ли сделать по данным результатам лечения обоснованный вывод о том, что новое лекарство более эффективно, чем применявшиеся ранее методы лечения? Аналогичный вопрос логично задать, говоря о новых правилах поступления в вуз, о новом методе обучения, о пользе быстрой ходьбы, о преимуществах новой модели автомобиля или технологического процесса и т.д.

Процедура сопоставления высказанного предположения (гипотезы) с выборочными данными называется *проверкой гипотез*.

Задачи статистической проверки гипотез ставятся в следующем виде: относительно некоторой генеральной совокупности высказывается та или иная гипотеза H. Из генеральной совокупности извлекается выборка. Требуется указать правило, при помощи которого можно было бы по выборке решить вопрос о том, следует ли отклонить гипотезу H или принять её.

Следует отметить, что статистическими методами гипотезу *можно только опровергнуть или не опровергнуть*, но не доказать.

Например, для проверки утверждения автора (гипотеза H), что «в рукописи нет ошибок», рецензент прочел (изучил) несколько страниц рукописи. Если он обнаружил хотя бы одну ошибку, то гипотеза H отвергается, в противном случае – не отвергается, говорят, что «результат проверки с гипотезой согласуется».

Выдвинутая гипотеза может быть правильной или неправильной, поэтому возникает необходимость ее проверки.

3.2. Статистическая гипотеза. Статистический критерий

Под *статистической гипотезой* (или просто *гипотезой*) понимается всякое высказывание (предположение) о генеральной совокупности, проверяемое по выборке.

Статистические гипотезы делятся на гипотезы о параметрах распределения известного вида (это так называемые *параметрические* гипотезы) и гипотезы о виде неизвестного распределения (*непараметрические* гипотезы).

Одну из гипотез выделяют в качестве *основной* (или *нулевой*) и обозначают H_0, а другую, являющуюся логическим отрицанием H_0, т.е. противоположную H_0 – в качестве *конкурирующей* (или *альтернативной*) гипотезы и обозначают H_1.

Гипотезу, однозначно фиксирующую распределение наблюдений, называют *простой* (в ней речь идет об одном значении параметра), в противном случае – *сложной*.

Например, гипотеза H_0, состоящая в том, что математическое ожидание случайной величины X равно a_0, то есть $M_x = a_0$, является простой. В качестве альтернативной гипотезы можно рассматривать одну из следующих гипотез: $H_1 : M_x > a_0$ (сложная гипотеза), $H_1 : M_x < a_0$ (сложная), $H_1 : M_x \neq a_0$ (сложная) или $H_1 : M_x = a_1$ (простая гипотеза).

Имея две гипотезы H_0 и H_1, надо на основе выборки $x_1,\ x_2,\ldots,\ x_n$ принять либо основную гипотезу H_0, либо конкурирующую H_1.

Правило, по которому принимается решение принять или отклонить гипотезу H_0 (соответственно отклонить или принять гипотезу H_1), называется *статистическим критерием* (или просто *критерием*) проверки гипотезы H_0.

Проверку гипотез осуществляют на основании результатов выборки x_1, $x_2,\ldots,\ x_n$, из которых формируют функцию выборки $T_n = T(x_1,\ x_2,\ldots,\ x_n)$, называемой *статистикой критерия*.

Основной принцип проверки гипотез состоит в следующем. Множество возможных значений статистики критерия T_n разбивается на два непересекающихся подмножества: критическую область S, то есть область отклонения гипотезы и область $\overline{S}$ принятия этой гипотезы. Если фактически наблюдаемое значение статистики критерия (то есть значение критерия, вычисленное по выборке: $T_{\text{набл}} = T(x_1,\ x_2,\ldots,\ x_n)$) попадает в критическую область S, то основная гипотеза H_0 отклоняется и принимается альтернативная гипотеза H_1; если же $T_{\text{набл}}$ попадает в $\overline{S}$, то принимается H_0, а H_1 отклоняется.

При проверке гипотезы может быть принято неправильное решение, то есть могут быть допущены ошибки двух родов:

Ошибка первого рода состоит в том, что отвергается нулевая гипотеза H_0, когда на самом деле она верна.

Ошибка второго рода состоит в том, что отвергается альтернативная гипотеза H_1, когда она на самом деле верна.

Рассматриваемые случаи наглядно иллюстрирует следующая таблица:

Гипотеза H_0	Отвергается	Принимается
верна	ошибка 1-го рода	правильное решение
неверна	правильное решение	ошибка 2-го рода

Вероятность ошибки первого рода (обозначается через α) называется *уровнем значимости критерия*.

Очевидно, $\alpha = P(H_1|H_0)$. Чем меньше α, тем меньше вероятность отклонить верную гипотезу. Допустимую ошибку первого рода обычно задают заранее.

В одних случаях считается возможным пренебречь событиями, вероятность которых меньше 0,05 ($\alpha = 0{,}05$ означает, что в среднем в 5 случаях из 100 испытаний верная гипотеза будет отвергнута), в других случаях, когда речь идет, например, о разрушении сооружений, гибели судна и т.п., нельзя пренебречь обстоятельствами, которые могут появиться с вероятностью, равной 0,001.

Обычно для α используются стандартные значения: $\alpha = 0{,}05$; $\alpha = 0{,}01$; 0,005; 0,001.

Вероятность ошибки 2-го рода обозначается через β, то есть $\beta = P(H_0|H_1)$. Величину $1-\beta$, то есть вероятность недопущения ошибки второго рода (отвергнуть неверную гипотезу H_0, принять верную H_1) называется *мощностью критерия*.

Очевидно, $1-\beta = P(H_1|H_1) = P\left((x_1,\ x_2,\ldots,\ x_n) \in S|H_1\right)$.

Чем больше мощность критерия, тем вероятность ошибки 2-го рода меньше, что, конечно, желательно (как и уменьшение α).

Последствия ошибок 1-го, 2-го рода могут быть совершенно различными: в одних случаях надо минимизировать α, в других – β. Так, применительно к производству, к торговле, можно сказать, что α – риск поставщика (то есть забраковка по выборке всей партии изделий, удовлетворяющих стандарту), β – риск потребителя (то есть прием по выборке всей партии изделий, не удовлетворяющих стандарту); применительно к судебной системе, ошибка 1-го рода приводит к оправданию виновного, ошибка 2-го рода – к осуждению невиновного.

Отметим, что одновременное уменьшение ошибок 1-го и 2-го рода возможно лишь при увеличении объема выборок. Поэтому обычно при заданном уровне значимости α отыскивается критерий с наибольшей мощностью.

Методика проверки гипотез сводится к следующему:

1. Располагая выборкой $X_1, X_2, \ldots, X_n$, формируют нулевую гипотезу H_0 и альтернативную H_1.
2. В каждом конкретном случае подбирают статистику критерия $T_n = T(X_1, X_2, \ldots, X_n)$, обычно из нижеперечисленных: U – нормальное распределение, χ^2 – распределение хи-квадрат Пирсона, t – распределение Стьюдента.
3. По статистике критерия T_n и уровню значимости α определяют критическую область S (и $\overline{S}$). Для ее отыскания достаточно найти критическую точку $t_{кр}$, то есть границу, отделяющую S от $\overline{S}$.

Границы областей определяются соответственно из соотношений:
$P(T_n > t_{кр}) = \alpha$, для правосторонней критической области S (рис. 5);

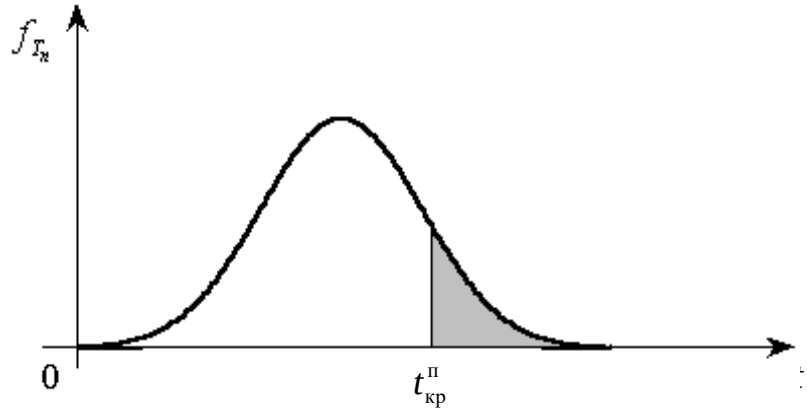

Рис. 5.Правосторонняя критическая область

$P(T_n < t_{кр}) = \alpha$, для левосторонней критической области S (рис. 6);

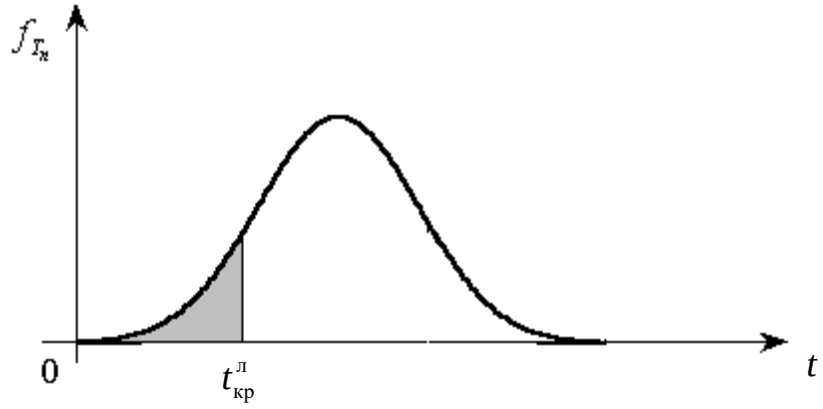

Рис. 6. Левосторонняя критическая область

$P\left(T_n < t_{\text{кр}}^{\text{л}}\right) = P\left(T_n > t_{\text{кр}}^{\text{п}}\right) = \dfrac{\alpha}{2}$, для двусторонней критической области S (рис. 7)

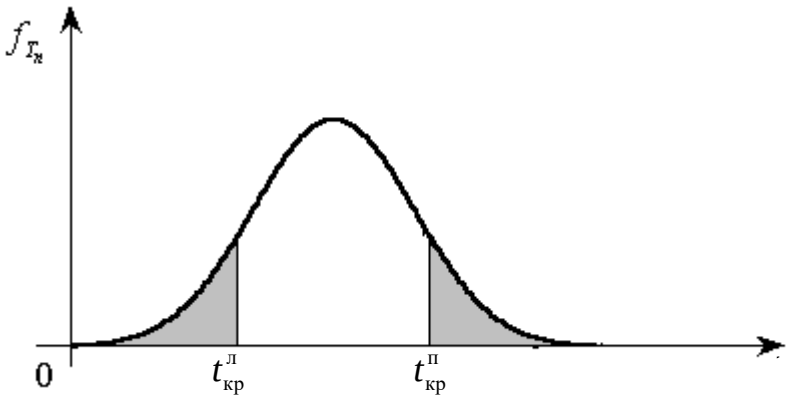

Рис. 7. Двусторонняя критическая область

Для каждого конкретного случая имеются соответствующие таблицы, по которым и находят критическую точку, удовлетворяющую приведенным выше соотношениям.

4. Для полученной реализации выборки $x = (x_1, x_2, ..., x_n)$ подсчитывают значение критерия, то есть $T_{\text{набл}} = T(x_1, x_2, ..., x_n) = t$.
5. Если $t \in S$ (например, $t > t_{\text{кр}}$ для правосторонней области S), то нулевую гипотезу H_0 отвергают; если же $t \in \overline{S}$ $(t < t_{\text{кр}})$, то нет оснований, чтобы отвергнуть гипотезу H_0.

3.3. Проверка гипотез о законе распределения

Во многих случаях закон распределения изучаемой случайной величины неизвестен, но есть основания предположить, что он имеет вполне определенный вид: нормальный, биномиальный или какой-либо другой.

Пусть необходимо проверить гипотезу H_0 о том, сто случайная величина подчиняется определенному закону распределения, заданному функцией распределения $F_0(x)$, то есть $H_0 : F_{\text{X}}(x) = F_0(x)$. Под альтернативной гипотезой будем понимать в данном случае то, что просто не выполнена основная (то есть $H_1 : F_{\text{X}}(x) \neq F_0(x)$).

Для проверки гипотезы о распределении случайной величины X проведем выборку, которую оформим в виде статистического ряда:

x_i	x_1	x_2	…	x_m
n_i	n_1	n_2	…	n_m

где $\sum_{i=1}^{m} n_i = n -$ объем выборки.

Требуется сделать заключение: согласуется ли результаты наблюдений с высказанным предположением. Для этого используем специально подобранную величину – критерий согласия.

Критерием согласия называют статистический критерий проверки гипотезы о предполагаемом законе неизвестного распределения. (он используется для проверки согласия предполагаемого вида распределения с опытными данными на основании выборки.)

Критерий согласия Пирсона – наиболее часто употребляемый критерий для проверки простой гипотезы о законе распределения.

Для проверки гипотезы H_0 поступают следующим образом.

Разбивают всю область значений случайной величины X на m интервалов $\Delta_1, \Delta_2, ..., \Delta_m$ и подсчитывают вероятности p_i $(i = 1, 2, ..., m)$ попадания случайной величины X (то есть наблюдения) в интервал Δ_i, используя формулу $P\{\alpha \le X \le \beta\} = F_0(\beta) - F_0(\alpha)$. Тогда теоретическое число значений случайной величины X, попавших в интервал Δ_i, можно рассчитать по формуле $n \cdot p_i$. Таким образом, получим теоретический ряд распределения:

Δ_i	Δ_1	Δ_2	…	Δ_m
$n_i' = np_i$	$n_1' = np_1$	$n_2' = np_2$	…	$n_m' = np_m$

Если эмпирические частоты (n_i) сильно отличаются от теоретических $(np_i = n_i')$, то проверяемую гипотезу H_0 следует отвергнуть, в противном случае принять.

Каким критерием, характеризующим степень расхождения между эмпирическими и теоретическими частостями, следует воспользоваться? В качестве меры расхождения между n_i и np_i для $i = 1, 2, ..., m$ К.Пирсон предложил величину:

$$\chi^2 = \sum_{i=1}^{m} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{m} \frac{n_i^2}{np_i} - n.$$

Согласно теореме Пирсона, при $n \to \infty$ статистика имеет χ^2-распределение с $k = m - r - 1$ степенями свободы , где $m -$ число групп (интервалов) выборки, $r -$ число параметров предполагаемого распределения. В частности, если предполагаемое распределение нормально, то оценивают два параметра $(a$ и $\sigma)$, поэтому число степеней свободы $k = m - 3$.

Правило применения критерия χ^2 сводится к следующему:

1. По формуле вычисляют $\chi^2_{\text{набл}}$ - выборочное значение статистики критерия.

2. Выбрав уровень значимости α, по таблице χ^2- распределения находим критическую точку $\chi^2_{\alpha,k}$.

3. Если $\chi^2_{\text{набл}} \le \chi^2_{\alpha,k}$, то гипотеза H_0 не противоречит опытным данным; если $\chi^2_{\text{набл}} > \chi^2_{\alpha,k}$, то гипотеза H_0 отвергается.

Необходимым условием применения критерия Пирсона является наличие в каждом из интервалов не менее 5 наблюдений (то есть $n_i \ge 5$). Если в отдельных интервалах их меньше, то число интервалов надо уменьшить путем объединения (укрупнения) соседних интервалов.

Пример 14. Измерены 100 обработанных деталей; отклонения от заданного размера приведены в таблице:

$[x_i, x_{i+1})$	$[-3,-2)$	$[-2,-1)$	$[-1,0)$	$[0,1)$	$[1,2)$	$[2,3)$	$[3,4)$	$[4,5)$
n_i	3	10	15	24	25	13	7	3

Проверить при уровне значимости $\alpha = 0{,}01$ гипотезу H_0 о том, что отклонения от проектного размера подчиняется нормальному закону распределения.

Решение.

Число наблюдений в крайних интервалах меньше 5, поэтому объединим их с соседними. Получим следующий ряд распределения ($n = 100$):

$[x_i, x_{i+1})$	$[-3,-1)$	$[-1,0)$	$[0,1)$	$[1,2)$	$[2,3)$	$[3,5)$
n_i	13	15	24	25	13	10

Случайную величину – отклонение – обозначим через X. Для вычисления вероятностей p_i необходимо вычислить параметры, определяющие нормальный закон распределения (a и σ). Их оценки вычислим по выборке:

$$\bar{x} = \frac{1}{100}\cdot\left(-2\cdot 13 + (-0{,}5)\cdot 15 + 0{,}5\cdot 24 + 1{,}5\cdot 25 + 2{,}5\cdot 13 + 4\cdot 10\right) = 0{,}885 \approx 0{,}9,$$

$$D_{\text{в}} = \frac{1}{n}\sum_{i=1}^{k} x_i^2\cdot n_i - \left(\bar{x}_{\text{в}}\right)^2 = \frac{1}{100}\cdot\left(4\cdot 13 + 0{,}25\cdot 15 + 0{,}25\cdot 24 + 2{,}25\cdot 25 + 6{,}25\cdot 13 + 16\cdot 10\right) -$$

$$-\left(0{,}885\right)^2 \approx 2{,}809, \quad \sigma = \sqrt{D_{\text{в}}} = \sqrt{2{,}809} \approx 1{,}676 \approx 1{,}7.$$

Находим p_i $(i = \overline{1,6})$. Так как случайная величина X подчиненная нормальному закону с параметрами (a, σ) определена на интервале $(-\infty, +\infty)$, то крайние интервалы в ряде распределения заменяем, соответственно на

$(-\infty,-1)$ и $(3,+\infty)$. Тогда $p_1 = P\{-\infty < X < -1\} = \Phi_0\left(\frac{-1-0{,}9}{1{,}7}\right) - \Phi_0(-\infty) = \frac{1}{2} -$

$\Phi(1{,}12) = 0{,}1314$.

Аналогично получаем:

$p_2 = 0{,}1667, \quad p_3 = 0{,}2258, \quad p_4 = 0{,}2183, \quad p_5 = 0{,}1503,$

$$p_6 = P\{3 \le X < \infty\} = \Phi_0(\infty) - \Phi_0\left(\frac{3-0{,}9}{1{,}7}\right) = 0{,}5 - \Phi_0(1{,}24) = 0{,}1075.$$

Полученные результаты приведем в следующей таблице:

$[x_i, x_{i+1})$	$(-\infty,-1)$	$[-1, 0)$	$[0, 1)$	$[1, 2)$	$[2, 3)$	$[3,+\infty)$
n_i	13	15	24	25	13	10
$n' = np_i$	13,14	16,67	22,58	21,83	15,03	10,75

Вычисляем $\chi^2_{\text{набл}}$:

$$\chi^2_{\text{набл}} = \sum_{i=1}^{6} \frac{n_i^2}{np_i} - n = \left(\frac{13^2}{13{,}14} + \frac{15^2}{16{,}67} + \frac{24^2}{22{,}58} + \frac{25^2}{21{,}83} + \frac{13^2}{15{,}03} + \frac{10^2}{10{,}75}\right) - 100 =$$

$$= 101{,}045 - 100 = 1{,}045,$$

$$\chi^2_{\text{набл}} = 1{,}045.$$

Находим число степеней свободы; по выборке рассчитаны два параметра, значит, $r = 2$. Количество интервалов 6, то есть $m = 6$. Следовательно, $k = 6 - 2 - 1 = 3.$ зная, что $\alpha = 0{,}01$ и $k = 3$, по таблице χ^2- распределения находим $\chi^2_{\alpha,k} = 11{,}3.$

Таким образом, $\chi^2_{\text{набл}} < \chi^2_{\alpha,k}$, $1{,}045 < 11{,}3$, следовательно, нет оснований отвергнуть проверяемую гипотезу.

4. КОРРЕЛЯЦИОННО-РЕГРЕССИОННЫЙ АНАЛИЗ

4.1. Понятие о корреляционной и регрессионной связи

Различают два вида зависимостей между экономическими явлениями: функциональную и корреляционную (статистическую).

При *функциональной* зависимости каждому значению независимой переменной X соответствует вполне определенное значение зависимой переменной Y.

В экономике в большинстве случаев между переменными величинами существуют зависимости, когда каждому значению одной переменной соответствует не какое-то определенное, а множество возможных значений

другой переменной. Иначе говоря, каждому значению одной переменной соответствует определенное (условное) распределение другой переменной. Такая зависимость получила название *статистической* (или *стохастической*, *вероятностной*)

Статистическую зависимость называют *корреляционной*, если при изменении значений одной величины меняется среднее значение другой.

При сравнении функциональных и корреляционных зависимостей следует иметь в виду, что при функциональной зависимости, зная X, можно вычислить величину Y, а при корреляционной зависимости устанавливается лишь тенденция изменения Y при изменении X.

Статистические связи между переменными можно изучать методами корреляционного и регрессионного анализа. Основной задачей *регрессионного анализа* является установление формы и изучение зависимости между переменными. Основной задачей *корреляционного анализа* – выявление связи между случайными величинами и оценка тесноты связи.

4.2. Коэффициент корреляции

Для характеристики корреляционной зависимости между случайными величинами вводится понятие коэффициента корреляции r.

Коэффициент корреляции между двумя случайными величинами X и Y вычисляется по формуле:

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y},$$

где $\sigma_x = \sqrt{\overline{x^2} - (\bar{x}_в)^2}$, $\sigma_y = \sqrt{\overline{y^2} - (\bar{y}_в)^2}$ - средние квадратические отклонения случайных величин X и Y соответственно.

Отметим некоторые *свойства* коэффициента корреляции:

1. Если X и Y независимые случайные величины, то коэффициент корреляции равен нулю.
2. Коэффициент корреляции принимает значения на отрезке $[-1,1]$, то есть $-1 \le r \le 1$. В зависимости от того, насколько $|r|$ приближается к 1, в математической статистике различают (шкала Шеддока): связи нет $(r < 0{,}2)$, связь слабую $(0{,}2 \le r < 0{,}5)$, умеренную $(0{,}5 \le r < 0{,}75)$, тесную $(0{,}75 \le r \le 0{,}95)$ и очень тесную $(0{,}95 \le r < 1)$.
3. Если $|r| = 1$, то между случайными величинами X и Y имеет место функциональная, а именно линейная зависимость.
4. Коэффициент корреляции указывает на направление связи. Если $r > 0$, то связь прямая, если $r < 0$ отрицателен, то это свидетельствует о наличии обратной связи.

Квадрат коэффициента корреляции называется коэффициентом *детерминации*:

$$\eta = r^2.$$

Коэффициент детерминации η показывает, какая часть общей вариации Y обусловлена вариацией X.

Пример 15. С целью анализа влияния заработной платы на текучесть рабочей силы на пяти однотипных предприятиях проведены измерения уровня зарплаты (тыс.руб.) X и числа уволившихся за год рабочих Y:

X	3	4	5	5,5	6
Y	60	35	20	20	15

Определить степень влияния заработной платы на текучесть рабочей силы.

Решение.

Для определения тесноты связи вычислим коэффициент корреляции, для чего составим расчетную таблицу:

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	**2**	**3**	**4**	**5**	**6**
1	3	60	9	3600	180
2	4	35	16	1225	140
3	5	20	25	400	100
4	5,5	20	30,25	400	110
5	6	15	36	225	90
$\sum$	**23,5**	**150**	**116,25**	**5850**	**620**

Так как коэффициент корреляции рассчитывается по формуле

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y}, \text{ то:}$$

1. Найдем средние значения: $\bar{x}$ (сумма значений второго столбца, деленная на число строк:

$$\bar{x} = \frac{\sum_{i=1}^{5} x}{n} = \frac{23{,}5}{5} = 4{,}7;$$

среднее значение $\bar{y}$ (сумма значений третьего столбца, деленная на число строк):

$$\bar{y} = \frac{\sum_{i=1}^{5} y}{n} = \frac{150}{5} = 30;$$

среднее значение $\overline{x\,y}$ (среднее значение шестого столбца):

$$\overline{x\,y} = \frac{\sum_{i=1}^{5} \overline{x\,y}}{n} = \frac{620}{5} = 124.$$

2. Найдем средние квадратические отклонения σ_x и σ_y:

$$\sigma_x = \sqrt{\overline{x^2} - (\overline{x}_в)^2} = \sqrt{\frac{\sum_{i=1}^{5} x^2}{n} - \left(\frac{\sum_{i=1}^{5} x}{n}\right)^2} =$$

$$= \sqrt{\frac{116,25}{5} - \left(\frac{23,5}{5}\right)^2} = \sqrt{23,25 - (4,7)^2} = \sqrt{1,16} \approx 1,077,$$

где $\overline{x^2}$ рассчитывается как среднее значение четвертого столбца.

Аналогично $\sigma_y = \sqrt{\overline{y^2} - (\overline{y}_в)^2} = \sqrt{1170 - 30^3} = \sqrt{270} \approx 16,432$,

где $\overline{y^2}$ - среднее значение пятого столбца.

3. Подставляя найденные значения в формулу коэффициента корреляции, получим:

$$r = \frac{124 - 4,7 \cdot 30}{16,432 \cdot 1,077} = \frac{-17}{17,697} = -0,96.$$

Таким образом, можно сделать вывод, что связь между заработной платой и текучестью рабочей силы очень тесная и обратная, так как полученный коэффициент корреляции отрицательный. Это говорит о том, что чем меньше заработная плата (X), тем больше число уволившихся.

Выясним, какая часть вариации Y обусловлена вариацией X. Вычислим коэффициент детерминации:

$$\eta = r^2 = (-0,96)^2 = 0,92.$$

То есть вариации текучести рабочей силы (Y) на 92% обусловлена вариацией заработной платы (X).

4.3. Линейная парная регрессия.

После того, как с помощью корреляционного анализа выявлено наличие статистических связей между переменными и оценена степень тесноты, обычно переходят к математическому описанию вида зависимостей с использованием регрессионного анализа. Если коэффициент корреляции $r < 0,2$, то согласно шкале Шеддока связи между переменными нет, а следовательно не имеет смысла описывать модель связи.

Регрессионная модель представляет собой математическое выражение, связывающее случайные величины X и Y. *Уравнение регрессии* – это зависимость величины Y от X.

Часто встречающейся моделью зависимости является *линейная парная корреляция*. Вообще говоря, уравнение регрессии может описывать взаимосвязь не двух, а более переменных (то есть быть не парной, а множественной). Кроме того, связь между переменными далеко не всегда линейна.

В общем случае уравнение регрессии имеет вид:

$$Y = \varphi(X, \beta) + \varepsilon,$$

где β – параметры модели, ε – ошибка наблюдений.

Уравнение парной линейной регрессии выглядит следующим образом:

$$\hat{y} = a\,x + b,$$

где a и b - параметры уравнения линейной регрессии.

Для нахождения параметром применяют *метод наименьших квадратов*, согласно которому неизвестные a и b выбираются таким образом, чтобы сумма квадратов отклонений эмпирических средних значений от значений, найденных по уравнению регрессии была минимальной:

$$\sum_{i=1}^{n} \left(\hat{y}_i - \bar{y}_i\right)^2 \to \min.$$

Получим систему нормальных уравнений для нахождения искомых параметров:

$$\begin{cases} a\,n + b\sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} y_i; \\ a\sum\limits_{i=1}^{n} x_i + b\sum\limits_{i=1}^{n} x_i^2 = \sum\limits_{i=1}^{n} x_i y_i. \end{cases}$$

Разделив обе части уравнений на n, получим систему нормальных уравнений в виде:

$$\begin{cases} a + b\,\bar{x} = \bar{y}, \\ a\,\bar{x} + b\overline{x^2} = \overline{xy}. \end{cases}$$

Решая систему уравнений, найдем:

$$b = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\overline{x^2} - \bar{x}^2}.$$

зная, что $\overline{x^2} - \bar{x}^2 = \sigma_x^2$ и формулу для вычисления коэффициента корреляции можем записать:

$$b = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x^2} = r \cdot \frac{\sigma_y}{\sigma_x}.$$

Коэффициент b называется *коэффициентом регрессии*. Он показывает, на сколько единиц в среднем изменяется переменная Y при изменении X на одну единицу.

Замечание. Знак коэффициента регрессии указывает на направление связи: если $b>0$, связь прямая, если $b<0$ - обратная. Очевидно, что знаки коэффициентов корреляции и регрессии должны совпадать.

Решая систему относительно параметра a, получим:

$$a=\frac{1}{n}\sum_{i=1}^{n}y_i-b\frac{1}{n}\sum_{i=1}^{n}x_i=\bar{y}-b\bar{x}.$$

Для установления влияния на зависимую переменную независимой переменной, то есть для интерпретации модели используется коэффициент эластичности:

$$Э_x=b\frac{\bar{x}}{\bar{y}}.$$

Коэффициент эластичности показывает, на сколько процентов изменится Y при изменении X на 1 %.

Пример 16. В условиях предыдущей задачи найти уравнение линейной регрессии, выражающее зависимость между заработной платой рабочих и числом уволившихся.

Решение.

1. Для определения параметров a и b линии регрессии $\hat{y}=a\,x+b$ составим систему нормальных уравнений:

$$\begin{cases}a+b\,\bar{x}=\bar{y},\\ a\,\bar{x}+b\overline{x^2}=\overline{xy}.\end{cases}$$

2. Подставляя найденные в предыдущей задаче средние значения $\bar{x}=4{,}7$, $\bar{y}=30$, $\overline{x^2}=23{,}25$ $\overline{xy}=124$, получим:

$$\begin{cases}a+4{,}7b=30,\\ 4{,}7a+23{,}25b=124.\end{cases}$$

3. Решая эту систему, найдем $b=-14{,}65$; $a=98{,}85$. Тогда уравнение регрессии:

$$\hat{y}=98{,}85-14{,}65\,x.$$

Отрицательный коэффициент регрессии подтверждает то, что связь между заработной платой рабочих и текучестью кадров обратная. Вычислим коэффициент эластичности:

$$Э_x=b\frac{\bar{x}}{\bar{y}}=-14{,}65\cdot\frac{4{,}7}{30}\approx-2{,}3\%.$$

Полученный коэффициент свидетельствует о том, что при увеличении заработной платы на 1%, число увольняющихся в среднем сократиться на 2,3%.

Указания к выполнению РГР

Вариант работы выбирается согласно общему списку группы, помещенному в журнале:

№ в списке группы	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23	24	25	26
№ варианта	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6	7	8	9	10	1	2	3	4	5	6

РГР состоит из двух задач, связанных с математическими расчетами, построением таблиц и вычерчиванием графиков. Решение задач должно быть представлено подробно (поэтапно) с указанием формулировки условия, соответствующих формул и вычислений. Таблицы и графики следует изображать с соблюдением общих правил оформления табличного и графического материала.

Варианты заданий для РГР

Задание 1. По имеющимся данным требуется:

1. Построить статистический ряд распределения. изобразить получившийся ряд графически с помощью полигона или гистограммы. Найти функцию распределения, построить ее график.
2. Найти: выборочную среднюю, выборочную дисперсию, среднее квадратическое отклонение выборки, моду и медиану.
3. Проверить при уровне значимости $\alpha = 0{,}05$ гипотезу о соответствии имеющего статистического распределения нормальному закону.
4. Считая данные нормально распределенной случайной величиной найти:

а) точечную оценку математического ожидания изучаемой совокупности;

б) доверительный интервал для математического ожидания с доверительной вероятностью 0,95.

Вариант 1. Имеются данные о численности работающих в фирмах отрасли, человек:

88	67	48	63	83	81	87	79	82
102	51	90	85	45	74	76	46	84
118	62	69	36	73	52	79	75	82
80	99	86	82	77	89	105	60	78
111	70	73	94	57	108	58	91	46

Вариант 2. Имеются следующие данные о количестве произведенной продукции по предприятиям отрасли, тыс.руб.:

520	490	180	600	300	380	900	570	590	750
810	600	200	450	400	450	450	340	700	520
750	400	230	340	200	400	340	780	490	410
600	450	400	1000	310	600	260	490	590	650

Вариант 3. Имеются данные о стаже работы сотрудников предприятия, лет.

9	8	15	11	9	18	1	21	17	3
2	12	15	6	26	17	3	11	12	14
10	9	12	19	12	5	7	4	15	18
11	10	16	8	6	19	12	11	10	20

Вариант 4. Имеются данные о средненедельном объеме продаж торговых точек фирмы в тыс.руб.

104,9	57,1	43,0	62,9	62,1	125,9	61,5	100,3	75,8	96,1
76,9	100,4	149,2	45,6	84,2	65,3	76,0	39,3	33,3	46,2
76,1	22,1	96,4	63,1	86,4	79,8	2,1	58,6	103,6	72,8
26,2	91,7	84,1	94,2	83,0	79,0	98,1	79,3	117,9	17,6
147,2	113,4	52,4	75,8	124,5	26,9	115,6	84,7	124,9	57,6

Вариант 5. Имеются данные о располагаемом денежном доходе на душу населения в руб. в 2006 г.

4276	3459	6789	6298	8430	3670	4326	5678	4370	10115
5626	12422	10446	14640	5922	7559	9734	2426	10690	8492
11101	12552	10624	7309	8310	7468	12389	10088	8437	10813
14154	9505	10546	8999	7816	6716	5676	10368	7510	13158
8439	13582	9275	9655	7945	12587	8488	3929	8668	11965

Вариант 6. По данным обследования автомобилей некоторой марки имеются данные о расходе бензина в литрах на 100 км:

7,9	7,6	7,3	7,0	5,7	7,1	8,0	6,6	7,4	7,3
7,5	6,5	7,0	6,9	7,2	7,3	7,0	6,3	5,9	7,0
6,8	7,5	6,6	7,7	7,1	5,7	7,0	6,9	8,1	6,6
6,8	8,2	6,9	6,8	7,2	6,4	6,9	6,7	6,5	7,1
7,1	7,3	6,4	6,1	6,7	7,3	6,9	6,8	7,6	7,0

Вариант 7. По данным статистического наблюдения имеются данные о количестве пользователей, посетивших сайт одной из торговых компаний, чел. в час:

40	47	24	29	41	21	59	22	54	56
41	58	55	37	36	32	48	35	39	45
30	34	35	39	45	44	37	40	47	52
33	40	61	42	55	45	26	43	38	31
45	53	38	35	62	44	25	57	27	28

Вариант 8. Имеются данные о прибыли коммерческих банков региона, млн. руб.:

7,9	13,7	36,8	5,3	38,1	25,6	12,5	23,9	19,1	7,3
25,1	15,4	2,0	22,1	40,3	18,0	37,5	34,0	9,7	20,3
23,5	13,4	26,7	0,2	28,5	0,1	37,6	27,2	34,4	13,6
27,6	33,5	49,3	45,2	16,8	25,3	35,4	25,3	31,7	5,1

Вариант 9. Имеются данные за несколько лет об урожайности пшеницы в некоторой области, ц/га:

33,4	27,0	31,2	23,2	19,1	24,4	36,6	26,6	30,6
35,5	19,5	31,1	23,2	31,4	28,1	28,3	23,3	27,9
33,9	20,3	30,9	32,7	31,7	24,2	31,2	31,5	19,9
26,9	33,9	32,3	42,1	34,5	32,0	28,7	21,6	38,8
29,1	37,5	23,6	25,9	23,2	28,4	20,0	37,8	21,5

Вариант 10. Получены следующие данные о технике чтения первоклассников в марте, слов в минуту:

56	63	79	61	13	55	45	62	59	45	63
38	56	15	48	64	42	34	65	44	63	52
34	47	57	52	111	33	66	10	72	58	19
45	104	99	89	59	28	72	70	57	40	41
43	34	54	75	87	99	84	42	31	95	39

Задание 2. По приведенным ниже данным требуется:

1. Оценить степень зависимости между переменными;
2. Найти уравнение линейной регрессии;
3. Интерпретировать полученную модель, сделать выводы.

Вариант 1. По годовым отчетам промышленных предприятий получена следующая информация:

Среднесписочное число работников, чел.	Объем продукции, млн.руб.	Среднесписочное число работников, чел.	Объем продукции, млн.руб.
700	402	1425	1756
1100	792	1208	1014
1285	1116	1400	1440
705	435	900	720
1300	1281	1300	1086
1450	1756	1480	1809
800	510	1295	1125
1380	1392	895	648
825	540	1440	1716
1210	924	1180	881

Вариант 2. По группе грузовых автотранспортных предприятий города имеется следующая информация за отчетный год:

Грузооборот, млн.ткм	Сумма затрат на перевозки, тыс. руб.	Грузооборот, млн.ткм	Сумма затрат на перевозки, тыс. руб.
62	1550	47	1245
40	1080	24	724
38	1033	18	579
25	750	58	1444
15	472	44	1145
10	840	33	699
52	1310	32	889
27	804	20	612

Вариант 3. Рабочие фирмы по производству пластиковых окон характеризуются следующими показателями:

Стаж работы, лет	Месячная зарплата, тыс. руб.	Стаж работы, лет	Месячная зарплата, тыс. руб.	Стаж работы, лет	Месячная зарплата, тыс. руб.
0	1,40	5	1,60	25	2,50
0	1,50	12	1,75	25	3,00
20	2,20	1	1,55	1	1,50
9	1,85	1	1,50	10	2,65
6	1,50	26	2,80	26	2,80
25	2,40	25	2,80	22	2,90
14	1,80	17	1,80	3	1,50
0	1,50	18	1,70	2	1,25
13	1,85	18	3,00	7	1,65

Вариант 4. Имеются следующие данные по группе промышленных предприятий за отчетный год:

Объем продукции, млн. руб.	Прибыль, тыс. руб.	Объем продукции, млн. руб.	Прибыль, тыс. руб.
197,7	13,5	204,7	30,6
592,0	136,2	466,8	111,8
465,5	97,6	292,2	49,6
296,2	44,4	423,1	105,8
584,1	146,0	192,6	30,7
480,0	110,4	360,5	64,8
578,5	138,7	208,3	33,3

Вариант 5. По сотовым телефонам некоторой марки имеются следующие данные:

Вес, гр.	Цена, усл.ед.	Вес, гр.	Цена, усл.ед.
92	320	87	146
115	970	106	460
90	112	93	335
70	87	90	437
73	97	65	1533
142	592	80	272
86	102	101	407
90	184	80	66
125	563	80	87
110	301	104	247
85	58	101	204

Вариант 6. Имеются следующие данные по промышленным заводам региона:

Основные производственные фонды, млн.руб.	Среднесписочное число работников чел.	Основные производственные фонды, млн.руб.	Среднесписочное число работников чел.	Основные производственные фонды, млн.руб.	Среднесписочное число работников чел.
13,3	280	56,6	990	18,0	430
21,1	480	63,0	930	22,0	510
28,0	503	31,0	560	10,0	340
38,0	710	28,0	610	16,0	390
55,0	1020	78,0	910	10,0	250
18,0	490	42,0	740	21,0	960
19,0	500	14,0	420	17,0	490
43,0	620	15,0	390	15,0	400

Вариант 7. Имеются следующие данные по предприятиям отрасли:

Основные производственные фонды, млн.руб.	Прибыль предприятия, тыс.руб.	Основные производственные фонды, млн.руб.	Прибыль предприятия, тыс.руб.	Основные производственные фонды, млн.руб.	Прибыль предприятия, тыс.руб.
31,1	340	23,0	200	62,0	600
42,0	400	26,0	230	39,0	400
28,0	230	38,5	490	29,0	300
55,0	500	78,0	380	60,0	600
61,2	600	42,0	450	41,8	450
65,0	800	44,0	400	27,0	340

23,0	200	57,0	600	88,0	1000
27,9	300	32,2	340	58,5	590
42,7	500	30,0	310	70,0	890
30,5	340	30,6	300	45,0	400
24,9	200	36,7	410	60,8	600
41,7	400	51,4	560	48,7	560
44,0	400	58,3	600	43,6	490

Вариант 8. Рабочие фирмы по производству металлических дверей характеризуются следующими показателями:

Стаж работы, лет	Выработка шт./чел.	Стаж работы, лет	Выработка, шт./чел.	Стаж работы, лет	Выработка, шт./чел.
0	28	5	48	25	60
0	35	12	50	25	70
20	68	20	65	25	62
20	65	1	42	1	40
9	55	1	40	10	65
20	65	2	42	16	54
6	45	26	70	26	65
25	68	25	70	22	68
14	55	17	60	3	32
0	40	18	55	2	20
13	56	18	71	7	43

Вариант 9. Имеются основные показатели деятельности коммерческих банков региона, млн. руб.:

Кредитные вложения, млн.руб.	Прибыль, млн. руб.	Кредитные вложения, млн.руб.	Прибыль, млн. руб.	Кредитные вложения, млн.руб.	Прибыль, млн. руб.
50,2	25,1	136,4	3,9	180,0	2,0
0,5	0,1	150,8	0,4	198,1	2,4
89,8	2,0	135,4	13,4	215,0	49,3
88,3	5,3	99,9	17,2	211,0	2,2
21,0	22,1	111,3	5,6	250,5	6,6
59,1	0,2	167,1	12,3	199,7	16,8
0,1	0,9	98,3	1,1	256,7	19,1
156,0	5,9	171,0	4,8	366,8	9,7
145,5	0,1	148,3	3,6	298,5	34,4
93,3	0,1	117,3	13,6	302,5	5,1

Вариант 10. Имеются следующие данные о работниках предприятия:

Стаж работы, лет	Месячная зарплата, руб.	Стаж работы, лет	Месячная зарплата, руб.	Стаж работы, лет	Месячная зарплата, руб.
10	800	5	300	10	700
4	690	6	450	18	850
6	690	7	500	6	650
9	700	3	700	10	750
3	560	18	890	8	500
9	700	5	720	9	800
12	670	19	810	17	700
20	1200	8	700	15	490
18	390	9	780	10	1350
16	600	5	590	10	820
2	590	6	500	8	900
5	600	7	500	5	600
7	480	2	760	8	800
9	700	12	900	9	730
12	750	16	920	10	630

Пример выполнения РГР

Задание 1. В результате статистического исследования, проведенного среди работников некоторого промышленного объединения на основе случайной выборки, получены следующие данные о величине совокупного месячного дохода (тыс. руб.)

1,806	10,530	7,540	5,347	4,601	6,115	2,189	6,992	5,479	3,829
4,519	3,188	4,496	7,868	5,201	7,337	7,293	4,439	2,712	3,283
6,370	3,324	4,891	3,830	3,782	5,703	6,135	4,537	8,074	3,942
8,025	4,685	3,749	4,582	6,580	5,430	6,252	6,928	6,508	6,377
7,602	3,852	5,564	4,005	3,954	4,185	4,324	0,502	5,958	4,869

Выполнить статистическую обработку полученных данных.

Решение.

1. Для полученной выборочной совокупности объемом $n = 50$:
а). Производим ранжирование выборочных данных.

0,502	1,806	2,189	2,712	3,188	3,283	3,324	3,749	3,782	3,829
3,830	3,852	3,942	3,954	4,005	4,185	4,324	4,439	4,496	4,519
4,537	4,582	4,601	4,685	4,869	4,891	5,201	5,347	5,430	5,479
5,564	5,703	5,958	6,115	6,135	6,252	6,370	6,377	6,508	6,580
6,928	6,992	7,293	7,540	7,602	7,868	8,025	8,074	8,337	10,530

б) Определяем минимальное и максимальное значение признака.

$$X_{\min} = 0{,}502 \text{ тыс.руб.}; \quad X_{\max} = 10{,}530 \text{ тыс.руб.}$$

в) Находим размах варьирования признака

$$R = X_{\max} - X_{\min} = 10{,}028 \text{ тыс.руб.}$$

г) Определяем число групп, на которые разбиваем выборочную совокупность (округление проводим до ближайшего целого)

$$k = 1 + 3{,}32 \cdot \lg n = 7.$$

д) Определяем длину интервала по формуле

$$h = R / k = 1{,}433$$

е) Определяем границы интервалов и группируем данные по соответствующим интервалам. Границы интервалов (a_i, b_i), $i=1,2,\ldots,k$, получаем следующим образом

$$a_1 = x_{\min}; \quad a_{i+1} = b_i = a_i + h + h; \quad b_k = x_{\max}.$$

Замечание. В данном случае за начало первого интервала принимаем $x_{\min}$, так как если воспользоваться формулой $x_{\text{нач}} = x_{\min} - \frac{h}{2}$, то получим $x_{\text{нач}} = 0{,}502 - \frac{1{,}433}{2} = -0{,}215$, что не имеет экономического смысла, то есть при определении границ интервала не стоит забывать об экономическом содержании задачи.

В процессе группировки определяем количество вариант, удовлетворяющих неравенствам $a_i < x \le b_i$, и строим интервальный вариационный ряд путем заполнения таблицы:

№ интервала	Границы интервала $a_i - b_i$	Частота m_i	Накопленная частота $m_{\sum i}$
0	1	3	4
1	0,502-1,935	2	2
2	1,935-3,367	5	7
3	3,367-4,800	17	24
4	4,800-6,232	11	35
5	6,232-7,665	10	45
6	7,665-9,097	4	49
$k = 7$	9,097-10,530	1	50
$\sum$	–	50	–

ж) На основе полученных данных строим статистический ряд распределения и его геометрические представления.

В пределах каждого интервала все значения признака приравниваем к его серединному значению $(a_i + b_i)/2$ и считаем, что частота относится именно к этому значению. Необходимые вычисления производим в таблице:

№ Интервала	Интервалы a_i - b_i	$x_i = \frac{(a_i + b_i)}{2}$	Частости $w_i = m_i / n$	Накопленные частости $w_{\sum i} = m_{\sum i} / n$	Относительная плотность распределения w_i / h
0	1	2	3	4	5
1	0,502-1,935	1,218	0,04	0,04	0,028
2	1,935-3,367	2,651	0,1	0,14	0,070
3	3,367-4,800	4,083	0,34	0,48	0,237
4	4,800-6,232	5,516	0,22	0,7	0,154
5	6,232-7,665	6,949	0,2	0,9	0,140
6	7,665-9,097	8,381	0,08	0,98	0,056
$k = 7$	9,097-10,530	9,814	0,02	1	0,014
$\sum$	—	—	1,00	—	—

Статистический ряд распределения образуют данные 2-го и 3-го столбцов таблицы. Для построения гистограммы распределения используются данные 1-го и 5-го столбцов, полигона -2-го и 5-го столбцов, кумуляты (функции распределения)– данные 1-го и 4-го столбцов.

Напомним, что для построения гистограммы по оси абсцисс откладываются частичные интервалы (a_i, b_i), на каждом из которых строим прямоугольник высотой w_i / h. Площадь ступенчатой фигуры, образуемой гистограммой, равна единице. Соединяя середины верхних оснований прямоугольников отрезками прямой, из гистограммы можно получить полигон распределения (рис. 8).

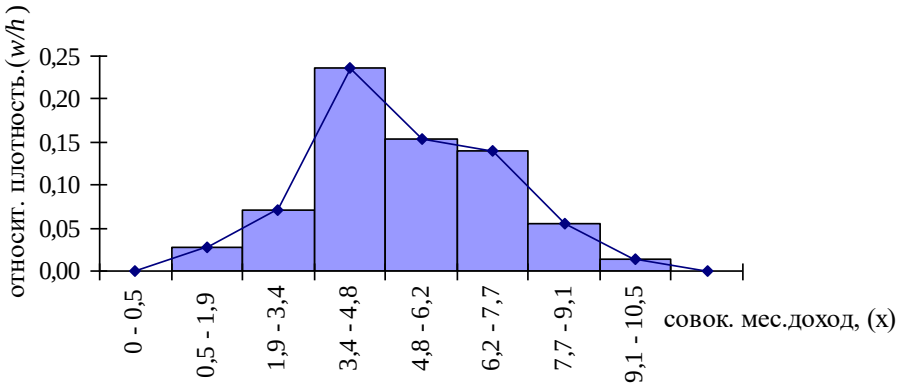

Рис. 8. Гистограмма и полигон распределения

При построении кумуляты в точках, соответствующих правому концу интервалов, по оси ординат откладываются накопленные частности $m_{\sum i}$, которые затем соединяются ломаной линией (рис.9)

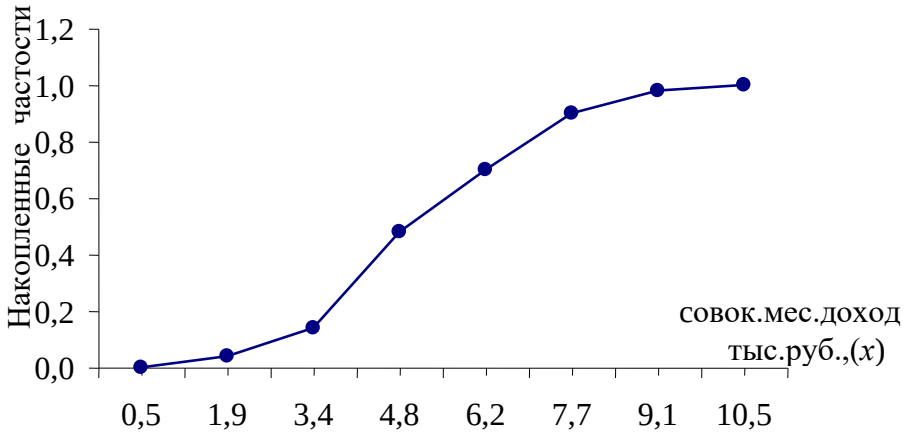

Рис 9. Кумулята распределения

2. Найдем выборочную среднюю, выборочную дисперсию, среднее квадратическое отклонение выборки, моду и медиану.

а) Вначале находим выборочное среднее, характеризующее центр распределения, около которого группируются выборочные данные, как взвешенное среднее

$$\bar{x} = \sum_{i=1}^{k} x_i w_i = 5{,}172 \text{ тыс.руб.}$$

Обозначая далее $\Delta_i = x_i - \bar{x}$, где $x_i = \frac{b_i - a_i}{2}$, вычисляем отклонения Δ_i варианты x_i от среднего значения $\bar{x}$ и заполняем таблицу:

№ п./п.	x_i	w_i	Δ_i	$x_i\, w_i$	$\Delta_i\, w_i$	$\Delta_i^2 w_i$
0	1	2		3	4	5
1	1,218	0,04	-3,954	0,049	-0,158	0,625
2	2,651	0,10	-2,521	0,265	-0,252	0,636
3	4,083	0,34	-1,089	1,388	-0,370	0,403
4	5,516	0,22	0,344	1,214	0,076	0,026
5	6,949	0,20	1,776	1,390	0,355	0,631
6	8,381	0,08	3,209	0,670	0,257	0,824
$k = 7$	9,814	0,02	4,642	0,196	0,093	0,431
$\sum$	–	1,00	–	5,172	0,000	3,576

Дисперсия выборочного распределения: $D_{\text{в}} = \sum_{i=1}^{7} \Delta_i^2 w_i = 3{,}576.$

Среднее квадратическое отклонение $\sigma_{\text{в}} = \sqrt{D_{\text{в}}} = \sqrt{3{,}576} \approx 1{,}89.$

В данном распределении модальным является интервал (3,367 – 4,800), так как ему соответствует наибольшая частота ($f = 17$). Значение моды определим по формуле:

$$Mo = x_{Mo} + h \cdot \frac{f_{Mo} - f_{(Mo-1)}}{\left(f_{Mo} - f_{(Mo-1)}\right) + \left(f_{Mo} + f_{(Mo-1)}\right)} =$$

$$= 3{,}367 + 1{,}433 \cdot \frac{17 - 5}{(17 - 5) + (17 - 11)} = 4{,}322.$$

Место медианы $N_{Me} = \frac{n+1}{2} = \frac{50+1}{2} = 25{,}5$, поэтому медианным является интервал (4,800 – 6,232), так как в этом интервале находятся номера 25 и 26. Вычислим медиану:

$$Me = x_{Me} + h \cdot \frac{\frac{\sum f}{2} - S_{(Me-1)}}{f_{Me}} = 4{,}800 + 1{,}433 \cdot \frac{25 - 24}{11} = 4{,}995.$$

3. Проверим гипотезу о соответствии имеющего статистического распределения нормальному закону.

Число наблюдений в крайних интервалах меньше 5, поэтому объединяем их с соседними. Получим:

Интервал	0,502 – 3,367	3,367 – 4,800	4,800 – 6,232	6,232 – 7,665	7,665 – 10,530
Частота , n_i	7	17	11	10	5

Оценки параметров распределения вычислим по выборке:

$a = \overline{x_{\text{в}}} \approx 5{,}2\,;\ \sigma = \sqrt{D_{\text{в}}} \approx 1{,}9$

$$p_i = p(a_i < X \le b_i) = \left[\Phi(t_{2i}) - \Phi(t_{1i})\right],$$

где $\Phi(t) = \frac{1}{\sqrt{2\pi}} \int_0^t e^{-x^2/2} dx$, $\quad t_{1i} = \frac{a_i - \bar{x}}{\sigma}$, $\quad t_{2i} = \frac{b_i - \bar{x}}{\sigma}$.

Плотность распределения вероятностей теоретического распределения на каждом интервале $(a_i - b_i)$ рассчитывается по формуле $f_i = p_i / h$.

Расчеты выполним в табличной форме:

№ п./п.	Интервалы a_i-b_i	w_i	t_{1i}	t_{2i}	$\Phi(t_{1i})$	$\Phi(t_{2i})$	p_i	np_i	n_i^2 / np_i
0	1	2	3	4	5	6	7	8	9
1	0,502-3,367	0,14	$-\infty$	-0,96	-0,500	-0,3315	0,1685	8,425	5,82
2	3,367-4,800	0,34	-0,96	-0,21	-0,3315	-0,0832	0,2483	12,415	23,28
3	4,800-6,232	0,22	-0,21	0,54	-0,0832	0,2054	0,2886	14,43	8,39
4	6,232-7,665	0,20	0,54	1,30	0,2054	0,4032	0,1978	9,89	10,11
5	7,665-10,530	0,10	1,30	∞	0,4032	0,500	0,0968	4,84	5,17
$\sum$	–	1,00	–	–	–	–	1,000	50	52,76

Вычисляем наблюдаемое значение критерия $\chi^2_{\text{набл}}$:

$$\chi^2_{\text{набл}} = \sum_{i=1}^{5} \frac{n_i^2}{np_i} - n = 52{,}76 - 50 = 2{,}76.$$

Число степеней свободы по выборке равно $k = m - r - 1$, где $m -$ число интервалов, $r -$ число параметров распределения, в нашем случае:

$$k = 5 - 2 - 1 = 2.$$

При уровне значимости $\alpha = 0{,}05$ и $k = 2$ по таблице распределения χ^2 находим $\chi^2_{\alpha,k} = 6{,}00$. Так как $\chi^2_{\text{набл}} < \chi^2_{\alpha,k}$ $(2{,}76 < 6{,}00)$, то нет оснований отвергнуть выдвинутую гипотезу.

4. Точечная оценка математического ожидания найдена при проверке гипотезы о соответствии распределения нормальному закону: $a = \overline{x_{\text{в}}} \approx 5{,}17$ (метод моментов).

Доверительный интервал для математического ожидания при известной дисперсии определяется из неравенства:

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

где t определяется из уравнения $\Phi_0(t) = \frac{\gamma}{2}$.

Учитывая, что $\gamma = 0{,}95$, получаем $\Phi_0(t) = 0{,}475$. По таблице находим $t = 1{,}96$. Тогда $t \cdot \frac{\sigma}{\sqrt{n}} = 1{,}96 \cdot \frac{1{,}9}{\sqrt{50}} \approx 0{,}53$. Доверительный интервал для математического ожидания будет:

$$5{,}17 - 0{,}53 < a < 5{,}17 + 0{,}53$$

$4{,}64 < a < 5{,}70$, то есть $(4{,}64; 5{,}70)$.

Задание 2. Имеются следующие данные о расходе бензина автомобилями некоторой марки:

Мощность двигателя, л.с.	Расход бензина, л. /100км.	Мощность двигателя, л.с.	Расход бензина, л. /100км.
110	9,5	143	12,7
87	5,7	150	13,0
115	9,0	205	18,0
90	6,1	125	11,0
160	14,5	140	12,6
84	6,0	200	17,5
190	17,4	175	16,1
105	8,0	220	19,7
99	7,3	132	10,2
240	22,0	165	15,0

Требуется:
1. Оценить степень зависимости между переменными;
2. Найти уравнение линейной регрессии;
3. Интерпретировать полученную модель, сделать выводы.

Решение.

1. Для определения тесноты связи вычислим коэффициент корреляции, для чего составим расчетную таблицу:

i	x_i	y_i	x_i^2	y_i^2	$x_i y_i$
1	**2**	**3**	**4**	**5**	**6**
1	84	6,0	7056	36,00	504,0
2	87	5,7	7569	32,49	495,9
3	90	6,1	8100	37,21	549,0
4	99	7,3	9801	53,29	722,7
5	105	8,0	11025	64,00	840,0
6	110	9,5	12100	90,25	1045,0
7	115	9,0	13225	81,00	1035,0
8	125	11,0	15625	121,00	1375,0
9	132	10,2	17424	104,04	1346,4
10	140	12,6	19600	158,76	1764,0
11	143	12,7	20449	161,29	1816,1
12	150	13,0	22500	169,00	1950,0
13	160	14,5	25600	210,25	2320,0
14	165	15,0	27225	225,00	2475,0
15	175	16,1	30625	259,21	2817,5
16	190	17,4	36100	302,76	3306,0
17	200	17,5	40000	306,25	3500,0
18	205	18,0	42025	324,00	3690,0
19	220	19,7	48400	388,09	4334,0
20	240	22,0	57600	484,00	5280,0
$\sum$	**2935**	**251,3**	**472049**	**3607,89**	**41165,6**

Коэффициент корреляции рассчитывается по формуле:

$$r = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y}.$$

а) Найдем средние значения:

$\bar{x}$ (сумма значений второго столбца, деленная на число строк:

$$\bar{x} = \frac{\sum_{i=1}^{20} x}{n} = \frac{2935}{20} = 146{,}75;$$

$\bar{y}$ (сумма значений третьего столбца, деленная на число строк):

$$\bar{y} = \frac{\sum_{i=1}^{20} y}{n} = \frac{251{,}3}{20} = 12{,}565;$$

$\overline{xy}$ (среднее значение шестого столбца):

$$\overline{xy} = \frac{\sum_{i=1}^{20} \overline{xy}}{n} = \frac{41165{,}6}{20} = 2058{,}28.$$

б) Найдем средние квадратические отклонения σ_x и σ_y:

$$\sigma_x = \sqrt{\overline{x^2} - (\bar{x}_{\text{в}})^2} == \sqrt{\frac{472049}{20} - (146{,}75)^2} = \sqrt{23602{,}45 - 21535{,}56} = \sqrt{2066{,}89} \approx 45{,}46,$$

где $\overline{x^2}$ рассчитывается как среднее значение четвертого столбца.

Аналогично $\sigma_y = \sqrt{\overline{y^2} - (\bar{y}_{\text{в}})^2} = \sqrt{180{,}39 - 12{,}57^3} = \sqrt{22{,}39} \approx 4{,}73,$

где $\overline{y^2}$ - среднее значение пятого столбца.

в) Подставляя найденные значения в формулу коэффициента корреляции, получим:

$$r = \frac{2058{,}28 - 146{,}75 \cdot 12{,}57}{45{,}46 \cdot 4{,}73} = \frac{213{,}63}{215{,}03} = 0{,}99.$$

2. Найдем уравнение линейной регрессии.

а) Для определения параметров a и b линии регрессии $\hat{y} = a\,x + b$ составим систему нормальных уравнений:

$$\begin{cases} a + b\,\bar{x} = \bar{y}, \\ a\,\bar{x} + b\overline{x^2} = \overline{xy}. \end{cases}$$

б) Подставляя найденные в предыдущем пункте задачи средние значения $\bar{x} = 146{,}75$, $\bar{y} = 12{,}57$, $\overline{x^2} = 23602{,}45$, $\overline{xy} = 2958{,}28$, получим:

$$\begin{cases} a + 146{,}75b = 12{,}57, \\ 146{,}75a + 23602{,}45b = 2058{,}28. \end{cases}$$

в). Решая эту систему, найдем $b = 0{,}1$; $a = -2{,}11$. Тогда уравнение регрессии имеет вид:

$$\hat{y} = -2{,}11 + 0{,}1x.$$

3. Таким образом, можно сделать вывод, что связь между мощностью двигателя и расходом бензина прямая и очень тесная, так как полученный коэффициент корреляции положительный и очень близок к единице. Это говорит о том, что чем больше мощность двигателя (X), тем больше расход бензина (Y).

Выясним, какая часть вариации Y обусловлена вариацией X, для этого вычислим коэффициент детерминации:

$$\eta = r^2 = (0{,}99)^2 = 0{,}98.$$

То есть вариация расхода бензина (Y) на 98% обусловлена вариацией мощности двигателя (X).

Положительный коэффициент регрессии $b = 0{,}1$ подтверждает то, что связь между мощностью двигателя и расходом топлива прямая. Вычислим коэффициент эластичности:

$$Э_x = b\frac{\bar{x}}{\bar{y}} = 0,1 \cdot \frac{146{,}75}{12{,}57} \approx 1,17\%.$$

Полученный коэффициент свидетельствует о том, что при увеличении мощности двигателя на 1%, расход бензина в среднем увеличится на 1,17 %.

СПИСОК РЕКОМЕНДУЕМОЙ ЛИТЕРАТУРЫ

1. Гмурман В.Е. Руководство к решению задач по теории вероятностей и математической статистике. М.: Высшая школа, 1979 – 400 с.

2. Гмурман В.Е. Теория вероятностей и математическая статистика. М.: «Высшая школа». – 1977 – 368 с.

3. Горелова Г.В., Кацко И.А. Теория вероятностей и математическая статистика в примерах и задачах с применением EXCEL: Учебное пособие для вузов – Ростов н/Д: Феникс, 2005.

4. Письменный Д.Т. Конспект лекций по теории вероятностей, математической статистике и случайным процессам – М.: Айрис-пресс, 2006.

5. Баврин И.И. Высшая математика: Учебник для студентов естественно-научных специальностей – М.: Издательский центр «Академия», 2002.

6. Кремер Н.Ш. Теория вероятностей и математическая статистика: Учебное пособие для вузов – М.: Юнити-Дана, 2000.

7. Сборник задач по математике для втузов. Часть 3. Теория вероятностей и математическая статистика./Под ред. Ефимова А.В. М.: Наука, 1990.

Приложение 1

Таблица значений функции $\varphi(x)=\frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}$

	0	1	2	3	4	5	6	7	8
0,0	0,3989	3989	3989	3988	3986	3984	3982	3980	3977
0,1	3970	3965	3961	3956	3951	3945	3939	3932	3925
0,2	3910	3902	3894	3885	3876	3867	3857	3847	3836
0,3	3814	3802	3790	3778	3765	3752	3739	3726	3712
0,4	3683	3668	3652	3637	3621	3605	3589	3572	3555
0,5	3521	3503	3485	3467	3448	3429	3410	3391	3372
0,6	3332	3312	3292	3271	3251	3230	3209	3187	3166
0,7	3123	3101	3079	3056	3034	ЗОН	2989	2966	2943
0,8	2897	2874	2850	2827	2803	2780	2756	2732	2709
0,9	2661	2637	2613	2589	2565	2541	2516	2492	2468
1,0	2420	2396	2371	2347	2323	2299	2275	2251	2227
1,1	2179	2155	2131	2107	2083	2059	2036	2012	1989
1,2	1942	1919	1895	1872	1849	1826	1804	1781	1758
1,3	1714	1691	1669	1647	1626	1604	1582	1561	1539
1,4	1497	1476	1456	1435	1415	1394	1374	1354	1334
1,5	1295	1276	1257	1238	1219	1200	1182	1163	1145
1,6	1109	1092	1074	1057	1040	1023	1006	0989	0973
1,7	0940	0925	0909	0893	0878	0863	0848	0833	0818
1,8	0790	0775	0761	0748	0734	0721	0707	0694	0681
1,9	0656	0644	0632	0620	0608	0596	0584	0573	0562
2,0	0540	0529	0519	0508	0498	0488	0478	0468	0459
2.1	0440	0431	0422	0413	0404	0396	0387	0379	0371
2,2	0355	0347	0339	0332	0325	0317	0310	0303	0297
2,3	0283	0277	0270	0264	0258	0252	0246	0241	0235
2,4	0224	0219	0213	0208	0203	0198	0194	0189	0184
2,5	0175	0171	0167	0163	0158	0154	0151	0147	0143
2,6	0136	0132	0129	0126	0122	0119	0116	0113	0110
2,7	0104	0101	0099	0096	0093	0091	0088	0086	0084
2,8	0079	0077	0075	0073	0071	0069	0067	0065	0063
2,9	0060	0058	0056	0055	0053	0051	0050	0048	0047
3,0	0044	0043	0042	0040	0039	0038	0037	0036	0035
3,1	0033	0032	0031	0030	0029	0028	0027	0026	0025
3,2	0024	0023	0022	0022	0021	0020	0020	0019	0018
3,3	0017	0017	0016	0016	0015	0015	0014	0014	0013
3,4	0012	0012	0012	0011	0011	0010	0014	0010	0009
3,5	0009	0008	0008	0008	0008	0007	0007	0007	0007
3,6	0006	0006	0006	0005	0005	0005	0005	0005	0005
3,7	0004	0004	0004	0004	0004	0004	0003	0003	0003
3,8	0004	0004	0004	0004	0004	0004	0003	0003	0003
3,9	0002	0002	0002	0002	0002	0002	0002	0002	0001

Приложение 2

Таблица значений функции $\Phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_0^x e^{-t^2/2} dt$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,00	0,0000	0,44	0,1700	0,88	0,3106	1,32	0,4066
0,01	0,0040	0,45	0,1736	0,89	0,3133	1,33	0,4082
0,02	0,0080	0,46	0,1772	0,90	0,3159	1,34	0,4099
0,03	0,0120	0,47	0,1808	0,91	0,3186	1,35	0,4115
0,04	0,0160	0,48	0,1844	0,92	0,3212	1,36	0,4131
0,05	0,0199	0,49	0,1879	0,93	0,3238	1,37	0,4147
0,06	0,0239	0,50	0,1915	0,94	0,3264	1,38	0,4162
0,07	0,0279	0,51	0,1950	0,95	0,3289	1,39	0,4177
0,08	0,0319	0,52	0,1985	0,96	0,3315	1,40	0,4192
0,09	0,0359	0,53	0,2019	0,97	0,3340	1,41	0,4207
0,10	0,0398	0,54	0,2054	0,98	0,3365	1,42	0,4222
0,11	0,0438	0,55	0,2088	0,99	0,3389	1,43	0,4230
0,12	0,0478	0,56	0,2123	1,00	0,3413	1,44	0,4251
0,13	0,0517	0,57	0,2157	1,01	0,3438	1,45	0,4265
0,14	0,0557	0,58	0,2190	1,02	0,3461	1,46	0,4279
0,15	0,0596	0,59	0,2224	1,03	0,3485	1,47	0,4292
0,16	0,0636	0,60	0,2257	1,04	0,3508	1,48	0,4306
0,17	0,0675	0,61	0,2291	1,05	0,3531	1,49	0,4319
0,18	0,0714	0,62	0,2324	1,06	0,3554	1,50	0,4332
0,19	0,0753	0,63	0,2357	1,07	0,3577	1,51	0,4345
0,20	0,0793	0,64	0,2389	1,08	0,3599	1,52	0,4357
0,21	0,0832	0,65	0,2422	1,09	0,3621	1,53	0,4370
0,22	0,0871	0,66	0,2454	1,10	0,3643	1,54	0,4382
0,23	0,0910	0,67	0,2486	1,11	0,3665	1,55	0,4394
0,24	0,0948	0,68	0,2517	1,12	0,3686	1,56	0,4406
0,25	0,0987	0,69	0,2549	1,13	0,3708	1,57	0,4418
0,26	0,1026	0,70	0,2580	1,14	0,3729	1,58	0,4429
0,27	0,1064	0,71	0,2611	1,15	0,3749	1,59	0,4441
0,28	0,1103	0,72	0,2642	1,16	0,3770	1,60	0,4452
0,29	0,1141	0,73	0,2673	1,17	0,3790	1,61	0,4463
0,30	0,1179	0,74	0,2703	1,18	0,3810	1,62	0,4474

Продолжение приложения 2

Таблица значений функции $\Phi(x) = \frac{1}{\sqrt{2\pi}} \cdot \int_0^x e^{-t^2/2} dt$

x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$	x	$\Phi(x)$
0,31	0,1217	0,75	0,2734	1,19	0,3830	1,63	0,4484
0,32	0,1255	0,76	0,2764	1,20	0,3849	1,64	0,4495
0,33	0,1293	0,77	0,2794	1,21	0,3869	1,65	0,4505
0,34	0,1331	0,78	0,2823	1,22	0,3883	1,66	0,4515
0,35	0,1368	0,79	0,2852	1,23	0,3907	1,68	0,4525
0,36	0,1406	0,80	0,2881	1,24	0,3925	1,68	0,4535
0,37	0,1443	0,81	0,2910	1,25	0,3944	1,69	0,4545
0,38	0,1480	0,82	0.2939	1,26	0.3952	1,70	0,4554
0,39	0,1517	0,83	0,2967	1,27	0.3980	1,71	0,4564
0,40	0,1554	0,84	0,2995	1,28	0,3997	1,72	0,4573
0,41	0,1591	0,85	0,3023	1,29	0,4015	1,73	0,4582
0,42	0,1628	0,86	0,3051	1,30	0,4032	1,74	0,4591
0,43	0,1664	0,87	0,3078	1,31	0,4049	1,75	0,4599
1,76	0,4608	1,97	0,4756	2,36	0,4909	2,78	0,4973
1,77	0,4616	1,98	0,4761	2,38	0,4913	2,80	0,4974
1,78	0,4525	1,99	0,4767	2,40	0,4918	2,82	0,4976
1,79	0,4633	2,00	0,4772	2,42	0,4922	2,84	0,4977
1,80	0,4641	2,02	0,4783	2,44	0,4927	2,86	0,4979
1,81	0,4649	2,04	0,4793	2,46	0,4931	2,88	0,4980
1,82	0,4656	2,06	0,4803	2,48	0,4934	2,90	0,4981
1,83	0,4664	2,08	0,4812	2,50	0,4938	2,92	0,4982
1,84	0,4671	2,10	0,4821	2,52	0,4941	2,94	0,4984
1,85	0,4678	2,12	0,4830	2,54	0,4945	2,96	0,4985
1,86	0,4686	2,14	0,4838	2,56	0,4948	2,98	0,4985
1,87	0,4693	2,16	0,4846	2,58	0,4951	3,00	0,49865
1,83	0.4699	2,18	0,4854	2,60	0,4953	3,20	0,49931
1,89	0.4706	2,20	0,4861	2,62	0,4956	3,40	0,49966
1,90	0,4713	2,22	0,4868	2,64	0,4959	3,60	0,499841
1,91	0,4719	2,24	0,4875	2,66	0,4961	3,80	0,499928
1,92	0,4726	2,26	0,4881	2,68	0,4963	4,00	0,499968
1,93	0,4732	2,28	0,4887	2,70	0,4965	4,50	0,499997
1,94	0,4738	2,30	0,4893	2,72	0,4967	5,00	0,499997
1,95	0,4744	2,32	0,4898	2,74	0,4969		
1,96	0,475	2,34	0,4904	2,76	0,4971		

Приложение 3.

Квантили $\chi^2_{\alpha,k}$ распределения χ^2_k (k — число степеней свободы)

k	Уровень значимости α					
	0,01	0,025	0,05	0,95	0,975	0,99
1	6,6	5,0	3,8	0,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	16,8	14,4	12,6	1,64	1,24	0,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,26
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,5
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,3	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0

Приложение 4.

Квантили t - распределения Стьюдента (k — число степеней свободы)

k	Уровень значимости α (двусторонняя критическая область)					
	0,10	0,05	0,02	0,01	0,002	0,001
1	6,31	12,7	31,82	63,7	318,3	637,0
2	2,92	4,30	6,97	9,92	22,33	31,6
3	2,35	3,18	4,54	5,84	10,22	12,9
4	2,13	2,78	3,75	4,60	7,17	8,61
5	2,01	2,57	3,37	4,03	5,89	6,86
6	1,94	2,45	3,14	3,71	5,21	5,96
7	1,89	2,36	3,00	3,50	4,79	5,40
8	1,86	2,31	2,90	3,36	4,50	5,04
9	1,83	2,26	2,82	3,25	4,30	3,78 I
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,20	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32
13	1,77	2,16	2,65	3,01	3,85	4,22
14	1,76	2,14	2,62	2,98	3,79	4,14
15	1,75	2,13	2,60	2,95	3,73	4,07
16	1,75	2,12	2,58	2,92	3,69	4,01
17	1,74	2,11	2,57	2,90	3,65	3,96
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,73	2,09	2,53	2,85	3,55	3,85
21	1,72	2,08	2,52	2,83	3,53	3,82
22	1,72	2,07	2,51	2,82	3,51	3,79
23	1,71	2,07	2,50	2,81	3,49	3,77
24	1,71	2,06	2,49	2,80	3,47	3,74
25	1,71	2,06	2,49	2,79	3,45	3,72
26	1,71	2,05	2,48	2,78	3,44	3,71
27	1,71	2,05	2,47	2,77	3,42	3,69
28	1,70	2,05	2,46	2,76	3,40	3,66
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3,65
40	1,68	2,02	2,42	2,70	3,31	3,55
60	1,67	2,00	2,39	2,66	3,23	3,46
120	1,66	1,98	2,36	2,62	3,17	3,37
00	1,64	1,96	2,33	2,58	3,09	3,29