3.1. Статистические оценки параметров распределения

Основные задачи математической статистики

Математическая статистика изучает закономерности, которые имеют место в массовых совокупностях однородных объектов.

Основные **задачи** математической статистики:

1. Разработка методов сбора и группировки статистических сведений, полученных в результате наблюдений или в результате специально поставленных экспериментов.
2. Разработка методов анализа статистических данных в зависимости от целей исследования. Сюда относятся:

а) оценка неизвестной вероятности события; оценка неизвестной функции распределения; оценка параметров распределения, вид которого известен; оценка зависимости случайной величины от одной или нескольких случайных величин и др.;

б) проверка статистических гипотез о виде неизвестного распределения или о величине параметров распределения, вид которого известен.

Современная математическая статистика разрабатывает способы определения числа необходимых испытаний до начала исследования (планирование эксперимента), в ходе исследования (последовательный анализ) и решает многие другие задачи. Современную математическую статистику определяют как науку о принятии решений в условиях неопределенности.

Итак, задача математической статистики состоит в создании методов сбора и обработки статистических данных для получения научных и практических выводов.

Генеральная и выборная совокупности. Виды выборки. Способы отбора

Пусть требуется изучить совокупность однородных объектов относительно некоторого качественного или количественного признака, характеризующего эти объекты.

Иногда проводят сплошное обследование, т.е. обследуют каждый из объектов совокупности относительно признака, которым интересуются. На практике, однако, сплошное обследование применяют сравнительно редко. Чаще случайно отбирают из всей совокупности ограниченное число объектов и подвергают их изучению.

Генеральной совокупностью называют совокупность однородных объектов, подлежащих изучению.

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

Объемом совокупности (выборной или генеральной) называют число объектов этой совокупности. Например, если из 1 000 деталей отобрано для обследования 100 деталей, то объем генеральной совокупности N=1 000, а объем выборки n=100.

При составлении выборки можно поступать двумя способами: после того как объект отобран и над ним произведено наблюдение, он может быть возвращен либо не возвращен в генеральную совокупность. В соответствии со сказанным, выборки подразделяют на повторные и безповторные.

Повторной называют выборку, при которой отобранный объект (перед отбором следующего) возвращается в генеральную совокупность.

Бесповторной называют выборку, при которой отобранный объект в генеральную совокупность не возвращается.

На практике обычно пользуются бесповторным случайным отбором.

Для того, чтобы по данным выборки можно было достаточно уверенно судить об интересующем признаке генеральной совокупности, необходимо, чтобы объекты выборки правильно его представляли. Другими словами, выборка должна правильно представлять пропорции генеральной совокупности. Это требование формулируют так: выборка должна быть ***репрезентативной (представительной).***

На практике применяют различные ***способы отбора.*** Принципиально эти способы можно подразделить на два вида:

1. Отбор, не требующий расчленения генеральной совокупность на части. Сюда относятся: а) простой случайный бесповторный отбор; б) простой случайный повторный отбор.
2. Отбор, при котором генеральная совокупность разбивается на части. Сюда относятся: а) типический отбор; б) механический отбор; в) серийный отбор.

Простым случайным называют такой отбор, при котором объекты извлекают по одному из всей генеральной совокупности. Осуществить простой отбор можно с помощью повторной и бесповторной выборки.

Типическим называют отбор, при котором объекты отбираются не из всей генеральной совокупности, а из каждой ее «типической» части. Например, если детали изготовляют на нескольких станках, то отбор производят не из всей совокупности деталей, произведенных всеми станками, а из продукции каждого станка в отдельности. Типическим отбором пользуются тогда, когда обследуемый признак заметно колеблется в различных типических частях генеральной совокупности. Например, если продукция изготовляется на нескольких машинах, среди которых есть более или менее изношенные, то здесь типический отбор целесообразен.

Механическим называют отбор, при котором генеральную совокупность «механически» делят на столько групп, сколько объектов должно войти в выборку, а из каждой группы отбирают один объект. Например, если нужно отобрать 20 % изготовленных станком деталей, то отбирают каждую пятую деталь; если требуется отобрать 5 % деталей, то отбирают каждую двадцатую деталь, и т.д.

Серийным называют отбор, при котором объекты отбирают из генеральной совокупности не по одному, а «сериями», которые подвергаются сплошному обследованию. Например, если изделия изготовляются большой группой станков-автоматов, то подвергают сплошному обследованию продукцию только нескольких станков. Серийным отбором пользуются тогда, когда обследуемый признак колеблется в различных сериях незначительно.

Подчеркнем, что на практике часто применяется комбинированный отбор, при котором сочетаются указанные выше способы. Например, иногда разбивают генеральную совокупность на серии одинакового объема, затем простым случайным отбором выбирают несколько серий и, наконец, из каждой серии простым случайным отбором извлекают отдельные объекты.

Статистическое распределение выборки

Пусть из генеральной совокупности извлечена выборка, причем x_1 наблюдалось n_1 раз, $x_2 - n_2$ раз, $x_k - n_k$ раз и $\sum n_i = n$ – объем выборки. Наблюдаемые значения x_i называют ***вариантами***, а последовательность вариант, записанных в возрастающем

порядке, – ***вариационным рядом***. Числа наблюдений называют ***частотами***, а их отношения к объему выборки $\frac{n_i}{n} = W_i$ – ***относительными частотами***.

Статистическим распределением выборки называют перечень вариант и соответствующих им частот или относительных частот.

Различают дискретные и интервальные статистические распределения.

Статистическое распределение называется ***дискретным***, если значения признака отличаются друг от друга не менее чем на некоторую постоянную величину

x_i	x_1	x_2	$\dots$	x_k
n_i	n_1	n_2	$\dots$	n_k
W_i	W_1	W_2	$\dots$	W_k

$$\sum_{i=1}^{k} n_i = n; \quad \sum_{i=1}^{k} W_i = 1.$$

Для графического представления дискретного распределения используют полигон частот (полигон относительных частот).

Полигоном частот называют ломаную, отрезки которой соединяют точки $(x_1; n_1), (x_2; n_2), \dots, (x_k; n_k)$. Для построения полигона на оси абсцисс откладывают варианты x_i, а на оси ординат – соответствующие им частоты n_i. Точки $(x_i; n_i)$ соединяют отрезками прямых и получают полигон частот (рис. 3.1).

Полигоном относительных частот называют, ломаную отрезки которой соединяют точки $(x_1; W_1), (x_2; W_2), \dots, (x_k; W_k)$. Для построения полигона относительных частот на оси абсцисс откладывают варианты x_i, а на оси ординат – соответствующие им относительные частоты W_i. Точки $(x_i; W_i)$ соединяют отрезками прямых и получают полигон относительных частот.

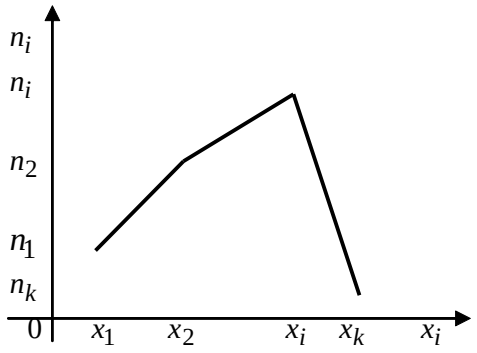


Рис. 3.1.

В случае непрерывных случайных величин рассматривают ***интервальное*** статистическое распределение выборки. Оно оформляется в виде следующей таблицы:

$(x_i; x_{i+1})$	$(x_1; x_2)$	$(x_2; x_3)$	$\dots$	$(x_{k-1}; x_k)$
n_i	n_1	n_2	$\dots$	n_{k-1}

W_i	W_1	W_2	$\cdots$	W_{k-1}

$$\sum_{i=1}^{k} n_i = n; \quad \sum_{i=1}^{k} W_i = 1.$$

Разница между двумя соседними вариантами называется ***шагом*** интервала $h = x_i - x_{i+1}$. От интервального распределения можно перейти к дискретному, взяв на каждом интервале $(x_i; x_{i+1})$ за отдельное значение x_i^* величину $x_i^* = \frac{x_i + x_{i+1}}{2}$, являющуюся серединой этого интервала.

Графической характеристикой интервальных распределений является гистограмма частот (гистограмма относительных частот).

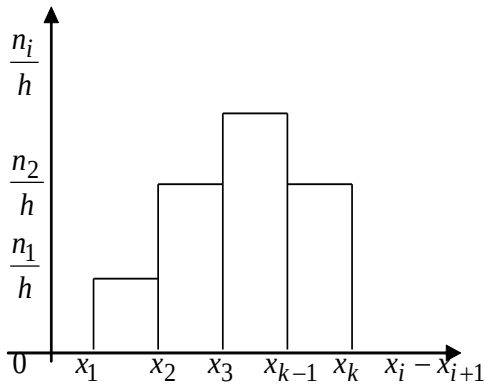


Рис. 3.2.

Гистограммой частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению $\frac{n_i}{h}$ (плотность частоты) (рис. 3.2).

Для построения гистограммы частот на оси абсцисс откладывают частичные интервалы, а над ними проводят отрезки, параллельные оси абсцисс, на расстоянии $\frac{n_i}{h}$.

Площадь *i*-го частичного прямоугольника равна $\frac{hn_i}{h} = n_i$ - сумме частот вариант *i*-го интервала; следовательно, ***площадь гистограммы частот равна сумме всех частот, т.е. объему выборки.***

Гистограммой относительных частот называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которых служат частичные интервалы длиною h, а высоты равны отношению $\frac{W_i}{h}$ (плотность относительной частоты).

Для построения гистограммы относительных частот на оси абсцисс откладывают частичные интервалы, а над ними проводят отрезки, параллельные оси абсцисс на расстоянии $\frac{W_i}{h}$.

Площадь *i*-го частичного прямоугольника равна $\frac{hW_i}{h} = W_i$ - относительной частоте вариант, попавших в *i*-й интервал. Следовательно, ***площадь гистограммы относительных частот равна сумме всех относительных частот, т.е. единице.***

По виду гистограмм можно предположить, какому теоретическому закону подчинен изучаемый признак генеральной совокупности. Форма гистограммы относительных частот дает представление о форме графика функции плотности $f(x)$ случайной величины.

Эмпирическая функция распределения

Пусть известно статистическое распределение частот количественного признака *X*. Введем обозначения: n_x - число наблюдений, при которых наблюдалось значение признака, меньшее *x*; *n* - общее число наблюдений (объем выборки). Ясно, что относительная частота события *X*<*x* равна $\frac{n_x}{n}$. Если *x* изменяется, то, вообще говоря, изменяется и относительная частота, т.е. относительная частота $\frac{n_x}{n}$ есть функция от *x*. Так как эта функция находится эмпирическим (опытным) путем, то ее называют эмпирической.

Эмпирической функцией распределения (функцией распределения выборки) называют функцию $F^*(x)$, определяющую для каждого значения *x* относительную частоту события *X*<*x*.

Итак, по определению,

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

где n_x - сумма частот вариант, меньших *x*; *n* – объем выборки.

Из определения функции $F^*(x)$ вытекают следующие ее свойства:

1. Значения эмпирической функции принадлежат отрезку $[0;1]$.
2. $F^*(x)$ - неубывающая функция.
3. Если x_1 - наименьшая варианта, то $F^*(x) = 0$ при $x \le x_1$; если x_k - наибольшая варианта, то $F^*(x) = 1$ при $x > x_k$.

Эмпирическая функция распределения выборки служит для оценки теоретической функции распределения генеральной совокупности.

Различие между эмпирической и теоретической функциями состоит в том, что теоретическая функция *F(x)* определяет вероятность события *X*<*x*, а эмпирическая функция $F^*(x)$ определяет относительную частоту этого же события.

Статистические оценки параметров распределения

Пусть требуется изучить количественный признак генеральной совокупности. Допустим, что из теоретических соображений удалось установить, какое именно распределение

имеет признак. Естественно возникает задача оценки параметров, которыми определяется это распределение. Например, если наперед известно, что изучаемый признак распределен в генеральной совокупности нормально, то необходимо оценить (приближенно найти) математическое ожидание и среднее квадратическое отклонение, так как эти два параметра полностью определяют нормальное распределение; если же есть основание считать, что признак имеет, например, распределение Пуассона, то необходимо оценить параметр λ, которым это распределение определяется.

Обычно в распоряжении исследователя имеются лишь данные выборки, например, значения количественного признака $x_1, x_2, \ldots, x_n$, полученные в результате n наблюдений. Через эти данные и выражают оцениваемый параметр. Рассматривая $x_1, x_2, \ldots, x_n$ как независимые случайные величины $X_1, X_2, \ldots, X_n$, можно найти статистическую оценку независимого параметра теоретического распределения.

Пусть одна из характеристик случайной величины X найдена приближенно, путем произведенных независимых опытов (испытаний), обозначим ее θ^*. Тогда случайная величина θ^* - ***статистическая оценка*** неизвестного параметра θ теоретического распределения количественного признака генеральной совокупности.

Статистическая оценка должна удовлетворять трем основным требованиям: несмещенности, эффективности и состоятельности.

Пусть произведено k опытов, в каждом из которых оценка θ^* приняла значения $\theta_1^*, \theta_2^*, \ldots, \theta_k^*$. Если оценка θ^* дает приближенное значение θ с избытком; тогда каждое найденное по данным выборок число $\theta_i^* (i = 1, 2, \ldots, k)$ больше истинного значения θ. Ясно, что в этом случае и математическое ожидание (среднее значение) случайной величины θ^* больше, чем θ, т.е. $M(\theta^*) > \theta$.

Таким образом, использование статистической оценки, математическое ожидание которой не равно оцениваемому параметру, привело бы к систематическим ошибкам. Требование $M(\theta^*) = \theta$ гарантирует избавление от этих ошибок.

Несмещенной называют статистическую оценку θ^*, математическое ожидание которой равно оцениваемому параметру θ при любом объеме выборки, т.е.

$$M(\theta^*) = \theta .$$

Смещенной называют оценку, математическое ожидание которой не равно оцениваемому параметру.

Однако было бы ошибочным считать, что несмещенная оценка всегда дает хорошее приближение оцениваемого параметра. Действительно, возможные значения θ^* могут быть сильно рассеяны вокруг своего среднего значения, т.е. дисперсия $D(\theta^*)$ может быть значительной. В этом случае найденная по данным одной выборки оценка, например θ_1^*, может оказаться весьма удаленной от среднего значения $\overline{\theta^*}$, а значит, и от самого оцениваемого параметра θ ; приняв θ_1^* в качестве приближенного значения θ, мы

допустили бы большую ошибку. Если же потребовать, чтобы дисперсия θ^* была малой, то возможность допустить большую ошибку будет исключена. По этой причине к статистической оценке предъявляется требование эффективности.

Эффективной называют статистическую оценку, которая (при заданном объеме выборки n) имеет наименьшую возможную дисперсию.

При рассмотрении выборок большего объема (n велико!) к статистическим оценкам предъявляется требование состоятельности.

Состоятельной называют статистическую оценку, которая при $n \to \infty$ стремится, по вероятности, к оцениваемому параметру, т.е. при увеличении количества опытов оценка θ^* параметра должна стремиться (сходиться) к истинному значению этого параметра.

Генеральная и выборочная средняя

Пусть изучается дискретная генеральная совокупность относительно количественного признака X.

Генеральной средней $\bar{x}_Г$ называют среднее арифметическое значений признака генеральной совокупности.

Если все значения $x_1, x_2, \ldots, x_N$ признака генеральной совокупности объема N различны, то

$$\bar{x}_Г = \frac{(x_1 + x_2 + \ldots + x_N)}{N}.$$

Если же значения признака $x_1, x_2, \ldots, x_k$ имеют соответственно частоты $N_1, N_2, \ldots, N_k$ причем $N_1 + N_2 + \ldots + N_k = N,$ то

$$\bar{x}_Г = \frac{(x_1 N_1 + x_2 N_2 + \ldots + x_k N_k)}{N},$$

т.е. генеральная средняя есть средняя взвешенная значений признака с весами, равными соответствующим частотам.

Замечание. Пусть генеральная совокупность объема N содержит объекты с различными значениями признака X, равными $x_1, x_2, \ldots, x_N$. Представим себе, что из этой совокупности наудачу извлекается один объект. Вероятность того, что будет извлечен объект со значением признака, например $x_1,$ очевидно, равна $\frac{1}{N}$. С этой же вероятностью может быть извлечен и любой другой объект. Таким образом, величину признака X можно рассматривать как случайную величину, возможные значения которой $x_1, x_2, \ldots, x_n$ имеют одинаковые вероятности, равные $\frac{1}{N}$. Найдем математическое ожидание $M(X)$:

$$M(X) = x_1 \cdot \frac{1}{N} + x_2 \cdot \frac{1}{N} + \ldots + x_N \cdot \frac{1}{N} = \frac{(x_1 + x_2 + \ldots + x_N)}{N} = \bar{x}_Г.$$

Итак, если рассматривать обследуемый признак X генеральной совокупности как случайную величину, то математическое ожидание признака равно генеральной средней этого признака:

$$M(X) = \bar{x}_Г.$$

Пусть для изучения генеральной совокупности относительно количественного признака X извлечена выборка объема n.

Выборочной средней $\bar{x}_B$ называют среднее арифметическое значение признака выборной совокупности.

Если все значения $x_1, x_2, \ldots, x_n$ признака выборки объема n различны, то

$$\bar{x}_B = \frac{(x_1 + x_2 + \ldots + x_n)}{n}.$$

Если же значения признака $x_1, x_2, \ldots, x_k$ имеют соответственно частоты $n_1, n_2, \ldots, n_k$, причем $n_1 + n_2 + \ldots + n_k = n$, то

$$\bar{x}_B = \frac{(x_1 n_1 + x_2 n_2 + \ldots + x_k n_k)}{n}$$

или

$$\bar{x}_B = \frac{\sum_{i=1}^{k} n_i x_i}{n},$$

т.е. выборочная средняя есть средняя взвешенная значений признака с весами, равными соответствующим частотам.

Оценка генеральной средней по выборочной средней

Пусть из генеральной совокупности извлечена повторная выборка объема n со значениями признака $x_1, x_2, \ldots, x_n$. Пусть генеральная средняя $\bar{x}_Г$ неизвестна и требуется оценить ее по данным выборки. В качестве оценки генеральной средней принимают выборочную среднюю

$$\bar{x}_B = \frac{(x_1 + x_2 + \ldots + x_n)}{n}.$$

Данная оценка удовлетворяет всем трем требованиям.

Докажем несмещенность, т.е.

$$M(\bar{x}_B) = \bar{x}_Г.$$

Будем рассматривать $\bar{x}_B$ как случайную величину и $x_1, x_2, \ldots, x_n$ как независимые, одинаково распределенные случайные величины $X_1, X_2, \ldots, X_n$. Поскольку эти величины одинаково распределены, то они имеют одинаковые числовые характеристики, в частности, одинаковое математическое ожидание, $M(x_i) = a$. Тогда

$$M(\bar{x}_B) = M\frac{(x_1 + x_2 + \ldots + x_n)}{n} = a.$$

С другой стороны, $M(X) = \bar{x}_Г = a$. В результате имеем $M(\bar{x}_B) = \bar{x}_Г$

Эффективность и состоятельность данной оценки предлагается доказать самостоятельно.

Генеральная и выборочная дисперсия

Для того чтобы охарактеризовать рассеяние значений количественного признака X генеральной совокупности вокруг своего среднего значения, вводят сводную характеристику – генеральную дисперсию.

Генеральной дисперсией $D_Г$ называют среднее арифметическое квадратов отклонений значений признака генеральной совокупности от их среднего значения $\bar{x}_Г$.

Если все значения $x_1, x_2, \ldots, x_N$ признака генеральной совокупности объема N различны, то

$$D_Г = \frac{\sum_{i=1}^{N}\left(x_i - \bar{x}_Г\right)^2}{N}.$$

Если же значения признака $x_1, x_2, \ldots, x_k$ имеют соответственно частоты $N_1, N_2, \ldots, N_k$ причем $N_1 + N_2 + \ldots + N_k = N,$ то

$$D_Г = \frac{\sum_{i=1}^{k} N_i\left(x_i - \bar{x}_Г\right)^2}{N},$$

т.е. генеральная дисперсия есть средняя взвешенная квадратов отклонений с весами, равными соответствующим частотам.

Выборочной дисперсией D_B называют среднее арифметическое квадратов отклонения наблюдаемых значений признака от их среднего значения $\bar{x}_B$.

Если все значения $x_1, x_2, \ldots, x_n$ признака выборки объема n различны, то

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

Если же значения признака $x_1, x_2, \ldots, x_k$ имеют соответственно частоты $n_1, n_2, \ldots, n_k$ причем $n_1 + n_2 + \ldots + n_k = n,$ то

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

т.е. выборочная дисперсия есть средняя взвешенная квадратов отклонений с весами, равными соответствующим частотам.

Формула для вычисления дисперсии

Вычисление дисперсии можно упростить, используя следующую теорему.

Теорема. Дисперсия равна среднему квадратов значений признака минус квадрат общей средней:

$$D = \overline{x^2} - \left(\bar{x}\right)^2.$$

Доказательство. Справедливость теоремы вытекает из преобразований:

$$D = \frac{\sum n_i\left(x_i - \bar{x}\right)^2}{n} = \frac{\sum n_i\left(x_i^2 - 2x_i\bar{x} + \left(\bar{x}\right)^2\right)}{n} =$$

$$= \frac{\sum n_i x_i^2}{n} - 2\bar{x}\frac{\sum n_i x_i}{n} + \left(\bar{x}\right)^2\frac{\sum n_i}{n} = \overline{x^2} - 2\bar{x}\cdot\bar{x} + \left(\bar{x}\right)^2 = \overline{x^2} - \left(\bar{x}\right)^2.$$

Итак, $D = \overline{x^2} - \left(\bar{x}\right)^2$, где $\bar{x} = \frac{\sum n_i x_i}{n}$, $\overline{x^2} = \frac{\sum n_i x_i^2}{n}$.

Оценка генеральной дисперсии по исправленной выборочной

Пусть из генеральной совокупности в результате n независимых наблюдений над количественным признаком X извлечена повторная выборка объема n

x_i	x_1	x_2	…	x_k
n_i	n_1	n_2	…	n_k

При этом $n_1 + n_2 + \ldots + n_k = n$.

Требуется по данным выборки оценить неизвестную генеральную дисперсию $D_Г$. Если в качестве оценки генеральной дисперсии принять выборочную дисперсию, то эта оценка будет приводить к систематическим ошибкам, давая заниженное значение генеральной дисперсии. Объясняется это тем, что, как можно доказать, выборочная дисперсия является смешенной оценкой $D_Г$, другими словами, математическое ожидание выборочной дисперсии не равно оцениваемой генеральной дисперсии, а равно

$$M[D_B] = \frac{n-1}{n} D_Г.$$

Легко «исправить» выборочную дисперсию так, чтобы ее математическое ожидание было равно генеральной дисперсии. Достаточно для этого умножить D_B на дробь $\frac{n}{(n-1)}$.

Сделав это, получим ***исправленную дисперсию***, которую обычно обозначают через s^2:

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

Исправленная дисперсия является, конечно, несмещенной оценкой генеральной дисперсии. Действительно,

$$M\left[s^2\right] = M\left[\frac{n}{n-1} D_B\right] = \frac{n}{n-1} M[D_B] = \frac{n}{n-1} \cdot \frac{n-1}{n} D_Г = D_Г.$$

Итак, в качестве оценки генеральной дисперсии принимают исправленную дисперсию

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

Для оценки же среднего квадратического отклонения генеральной совокупности используют «исправленное» среднее квадратическое отклонение, которое равно квадратному корню из исправленной дисперсии:

$$s = \sqrt{\frac{\sum_{i=1}^{k} n_i \left(x_i - \bar{x}_B\right)^2}{(n-1)}}.$$

Другие характеристики вариационного ряда

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

Модой M_0 называют варианту, которая имеет наибольшую частоту. Например, для ряда

варианта	1	4	7	9
частота	5	1	20	6

мода равна 7.

Медианой m_e называют варианту, которая делит вариационный ряд на две части, равные по числу вариант. Если число вариант нечетно, т.е. $n = 2k + 1$, то $m_e = x_{k+1}$; при четном $n = 2k$ медиана

$$m_e = \frac{\left(x_k + x_{k+1}\right)}{2}.$$

Например, для ряда 2 3 5 6 7 медиана равна 5; для ряда 2 3 5 6 7 9 медиана равна $\frac{(5+6)}{2} = 5{,}5.$

Размахом варьирования R называют разность между наибольшей и наименьшей вариантами:

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

Например, для ряда 1 3 4 5 6 10 размах равен 10-1=9.

Размах является простейшей характеристикой рассеяния вариационного ряда.

Средним абсолютным отклонением θ называют среднее арифметическое абсолютных отклонений:

$$\theta = \frac{\sum n_i \left| x_i - \bar{x}_B \right|}{\sum n_i}.$$

Например, для ряда

x_i	1	3	6	16
n_i	4	10	5	1

Имеем

$$\bar{x}_B = \frac{4 \cdot 1 + 10 \cdot 3 + 5 \cdot 6 + 1 \cdot 16}{4 + 10 + 5 + 1} = \frac{80}{20} = 4;$$

$$\theta = \frac{4 \cdot |1 - 4| + 10 \cdot |3 - 4| + 5 \cdot |6 - 4| + 1 \cdot |16 - 4|}{20} = 2{,}2.$$

Среднее абсолютное отклонение служит для характеристики рассеяния вариационного ряда.

Коэффициентом вариации V называют выраженное в процентах отношение выборочного среднего квадратического отклонения к выборочной средней:

$$V = \frac{\sigma_B}{\bar{x}_B} \cdot 100\%.$$

Коэффициент вариации служит для сравнения величин рассеяния по отношению к выборочной средней двух вариационных рядов: тот из рядов имеет большее рассеяние по

отношению к выборочной средней, у которого коэффициент вариации больше. Коэффициент вариации – безразмерная величина, поэтому он пригоден для сравнения рассеяний вариационных рядов, варианты которых имеют различную размерность.

Условные варианты

Предположим, что варианты выборки расположены в возрастающем порядке, т.е. в виде вариационного ряда.

Равноотстоящими называют варианты, которые образуют арифметическую прогрессию с разностью h.

Условными называют варианты, определяемые равенством

$$u_i = \frac{(x_i - C)}{h},$$

где C – ложный нуль; h – шаг, т.е. разность между любыми двумя соседними первоначальными вариантами.

Условные варианты используют для упрощенного вычисления числовых характеристик выборки.

Замечание. В качестве ложного нуля можно принять любую варианту. Максимальная простота вычислений достигается, если выбрать в качестве ложного нуля варианту, которая расположена примерно в середине вариационного ряда (часто такая варианта имеет наибольшую частоту). Варианте, которая принята в качестве ложного нуля, соответствует условная варианта, равная нулю.

Пример. Найти условные варианты статистического распределения:

варианта	23,6	28,6	33,6	38,6	43,6
частота	5	20	50	15	10

Решение. Выберем в качестве ложного нуля варианту 33,6 (эта варианта расположена в середине вариационного ряда).

Найдем шаг: $h = 28{,}6 - 23{,}6 = 5$.

Найдем условную варианту:

$$u_1 = \frac{(x_1 - C)}{h} = \frac{(23{,}6 - 33{,}6)}{5} = -2.$$

Аналогично получим: $u_2 = -1,\ u_3 = 0,\ u_4 = 1,\ u_5 = 2.$ Мы видим, что условные варианты – небольшие целые числа. Разумеется, оперировать с ними проще, чем с первоначальными вариантами.

Сведение первоначальных вариант к равноотстоящим

На практике, как правило, данные наблюдений не являются равноотстоящими числами. Для того чтобы свести выборку наблюдаемых значений признака к случаю равностоящих вариант существует следующий алгоритм:

1. Интервал, в котором заключены все наблюдаемые значения признака (первоначальные варианты), делят на несколько равных частных интервалов $(x_1; x_i), (x_i; x_j), \ldots, (x_k; x_m)$.

2. Находят середины частичных интервалов, которые и образуют последовательность равностоящих вариант

$$y_1 = \frac{x_1 + x_i}{2};\ y_2 = \frac{x_i + x_j}{2}; \ldots y_r = \frac{x_k + x_m}{2}.$$

В качестве частоты каждой «новой» варианты принимают общее число первоначальных вариант, попавших в соответствующий частичный интервал.

$$n_1' = n_1 + n_2 + \ldots + n_{i-1} + \frac{n_i}{2};$$

$$n_2' = \frac{n_i}{2} + n_{i+1} + \ldots + n_{j-1} + \frac{n_j}{2};$$

$$n_r' = \frac{n_k}{2} + n_{k+1} + \ldots + n_{m-1} + \frac{n_m}{2}.$$

Обычные, начальные, центральные, условные эмпирические моменты

Для вычисления сводных характеристик удобно пользоваться эмпирическими моментами, которые вычисляют по данным наблюдений.

Обычным эмпирическим моментом порядка k называют среднее значение k-х степеней разностей $x_i - C$:

$$M_k' = \frac{\sum n_i (x_i - C)^k}{n},$$

где x_i - наблюдаемая варианта; n_i - частота варианты; $n = \sum n_i$ - объем выборки; C – ложный нуль.

Начальным эмпирическим моментом порядка k называют обычный момент порядка k при C=0

$$M_k = \frac{\sum n_i x_i^k}{n}.$$

В частности,

$$M_1 = \frac{\sum n_i x_i}{n} = \bar{x}_B,$$

т.е. начальный эмпирический момент первого порядка равен выборочной средней.

Центральным эмпирическим моментом порядка k называют обычный момент порядка k при $C = \bar{x}_B$

$$m_k = \frac{\sum n_i (x_i - \bar{x}_B)^k}{n}.$$

В частности,

$$m_2 = \frac{\sum n_i (x_i - \bar{x}_B)^2}{n} = D_B,$$

т.е. центральный эмпирический момент второго порядка равен выборочной дисперсии.

Вычисление центральных моментов требует довольно громоздких вычислений. Чтобы упростить расчеты, заменяют первоначальные варианты условными.

Условным эмпирическим моментом порядка k называют начальный момент порядка *k*, вычисленный для условных вариант:

$$M_k^* = \frac{\sum n_i u_i^k}{n} = \frac{\sum n_i \left(\frac{x_i - C}{h} \right)^k}{n}.$$

В частности,

$$M_1^* = \frac{\sum n_i \left(\frac{x_i - C}{h} \right)}{n} = \frac{1}{h}\left[\frac{\sum n_i x_i}{n} - C \frac{\sum n_i}{n} \right] = \frac{1}{h}\left(\bar{x}_B - C \right).$$

Отсюда

$$\bar{x}_B = M_1^* h + C.$$

Таким образом, для того чтобы найти выборочную среднюю, достаточно вычислить условный момент первого порядка, умножить его на *h* и к результату прибавить ложный нуль *C*.

Для вычисления выборочной дисперсии можно воспользоваться формулой

$$D_B = \left[M_2^* - \left(M_1^* \right)^2 \right] \cdot h^2.$$

Действительно:

$$\left[M_2^* - \left(M_1^* \right)^2 \right] \cdot h^2 = \left[\frac{\sum n_i \left(\frac{x_i - C}{h} \right)^2}{n} - \left(\frac{\sum n_i \left(\frac{x_i - C}{h} \right)}{n} \right)^2 \right] \cdot h^2 =$$

$$= \frac{\sum n_i \overline{x_i^2}}{n} - 2C \frac{\sum n_i x_i}{n} + C^2 \frac{\sum n_i}{n} - \left(\frac{\sum n_i x_i}{n} \right)^2 + 2C \frac{\sum n_i x_i}{n} - C^2 \frac{\sum n_i}{n} =$$

$$= \overline{x^2} - \left(\bar{x} \right)^2 = D_B.$$

Метод произведений для вычисления выборочных средней и дисперсии

Цель метода заключается в нахождении условных эмпирических моментов и с их помощью D_B и $\bar{x}_B$.

Алгоритм метода

1. Составляется таблица, в первый столбец которой записывают выборочные варианты, располагая их в возрастающем порядке.
2. Во второй столбец записывают частоты вариант; и их сумму (объем выборки *n*) помещают в нижнюю клетку столбца.
3. В третий столбец записывают условные варианты $u_i = \frac{(x_i - C)}{h}$, причем в качестве ложного нуля *C* выбирают варианту, которая расположена примерно в середине вариационного ряда.

4. Умножают частоты на условные варианты и записывают их произведения $n_i u_i$ в четвертый столбец; сложив все полученные числа, их сумму $\sum n_i u_i$ помещают в нижнюю клетку столбца.

5. Умножают частоты на квадраты условных вариант и записывают их произведения $n_i u_i^2$ в пятый столбец; сложив все полученные числа, их сумму $\sum n_i u_i^2$ помещают в нижнюю клетку столбца.

6. Умножают частоты на квадраты условных вариант, увеличенных каждая на единицу, и записывают произведения $n_i (u_i + 1)^2$ в шестой контрольный столбец; сложив все полученные числа, их сумму $\sum n_i (u_i + 1)^2$ помещают в нижнюю клетку столбца.

7. На основе данных таблицы вычисляют условные моменты: первого и второго порядка:

$$M_1^* = \frac{\sum n_i u_i}{n},\ M_2^* = \frac{\sum n_i u_i^2}{n}.$$

8. Вычисляют выборочную среднюю и дисперсию по формулам

$$\bar{x}_B = M_1^* h + C,\ D_B = \left[M_2^* - \left(M_1^* \right)^2 \right] \cdot h^2.$$

Замечание. Шестой столбец служит для контроля вычислений: если сумма $\sum n_i (u_i + 1)^2$ окажется равной сумме $\sum n_i u_i^2 + 2\sum n_i u_i + n$ (как и должно быть в соответствии с тождеством $\sum n_i (u_i + 1)^2 = \sum n_i u_i^2 + 2\sum n_i u_i + n$), то вычисления проведены правильно.

Точность оценки, доверительная вероятность. Доверительный интервал

Точечной называют оценку, которая определяется одним числом. Все оценки, рассмотренные выше, точечные. При выборки малого объема точечная оценка может значительно отличаться от оцениваемого параметра, т.е. приводить к грубым ошибкам. По этой причине при наибольшем объеме выборки следует пользоваться интервальными оценками.

Интервальной называют оценку, которая определяется двумя числами – концами интервала. Интервальные оценки позволяют установить точность и надежность оценок.

Пусть найденная по данным выборки статистическая характеристика θ^* служит оценкой неизвестного параметра θ. Ясно, что θ^* тем точнее определяет параметр θ, чем больше абсолютная величина разности $\left|\theta - \theta^*\right|$. Другими словами, если $\delta > 0$ и $\left|\theta - \theta^*\right| < \delta$, то чем меньше δ, тем оценка точнее. Таким образом, положительное число δ характеризует ***точность оценки***.

Однако статистические методы не позволяют категорически утверждать, что оценка θ^* удовлетворяет неравенству $\left|\theta - \theta^*\right| < \delta$; можно лишь говорить о вероятности γ, с которой это неравенство осуществляется.

Надежностью (доверительной вероятностью) оценки θ по θ^* называют вероятность γ, с которой осуществляется неравенство $\left|\theta - \theta^*\right| < \delta$.

Обычно надежность оценки задается наперед, причем в качестве γ берут число, близкое к единице. Наиболее часто задают надежность, равную 0,95; 0,99 и 0,999.

Пусть вероятность того, что $\left|\theta - \theta^*\right| < \delta$, равна γ

$$P\left(\left|\theta - \theta^*\right| < \delta\right) = \gamma.$$

Заменив неравенство $\left|\theta - \theta^*\right| < \delta$ равносильным ему двойным неравенством $-\delta < \theta - \theta^* < \delta$ или $\theta^* - \delta < \theta < \theta^* + \delta$, имеем

$$P\left(\theta^* - \delta < \theta < \theta^* + \delta\right) = \gamma.$$

Это соотношение следует понимать так: вероятность того, что интервал $\left(\theta^* - \delta; \theta^* + \delta\right)$ заключает в себе (покрывает) неизвестный параметр θ, равно γ.

Доверительным называют интервал $\left(\theta^* - \delta; \theta^* + \delta\right)$, который покрывает неизвестный параметр с заданной надежностью γ.

Доверительный интервал для оценки математического ожидания нормального распределения при известном σ

Пусть количественный признак X генеральной совокупности распределен нормально, причем среднее квадратическое отклонение σ этого распределения известно. Требуется оценить неизвестное математическое ожидание a по выборочной средней $\bar{x}_B$. Найдем доверительные интервалы, покрывающие параметр a с надежностью γ.

Пусть дана выборочная совокупность объема n: $x_1, x_2, \ldots, x_n$. Так как количественный признак X генеральной совокупности распределен нормально, то элементы выборочной совокупности можно рассматривать как независимые случайные величины: $X_1, X_2, \ldots, X_n$ с одним и тем же законом распределения – нормальным, с параметрами $M(X_i) = a, \sigma(X_i) = \sigma$.

Воспользуемся доказательством следующей теоремы.

Теорема. Числовые характеристики среднего арифметического одинаково распределенных взаимно независимых случайных величин вычисляются по формулам

$$M(\overline{X}) = a;\ D(\overline{X}) = \frac{D}{n};\ \sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}.$$

Доказательство

$$M(\overline{X}) = M\,\frac{(X_1 + X_2 + \ldots + X_n)}{n} = \frac{M(X_1) + M(X_2) + \ldots + M(X_n)}{n} = \frac{na}{n} = a.$$

$$D(\overline{X}) = D\frac{(X_1 + X_2 + \ldots + X_n)}{n} = \frac{D(X_1) + D(X_2) + \ldots + D(X_n)}{n^2} = \frac{nD}{n^2} = \frac{D}{n}.$$

Очевидно, что $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$.

Потребуем, чтобы выполнялось соотношение

$$P\left(\left|\overline{X} - a\right| < \delta\right) = \gamma,$$

где γ - заданная надежность.

Для нормально распределенной случайной величины имеет место теорема

$$P\left(\left|X - a\right| < \delta\right) = 2\Phi\left(\frac{\delta}{\sigma}\right),$$

заменив X на $\overline{X}$ и σ на $\sigma(\overline{X}) = \frac{\sigma}{\sqrt{n}}$, получим

$$P\left(\left|\overline{X} - a\right| < \delta\right) = 2\Phi\left(\frac{\delta\sqrt{n}}{\sigma}\right) = 2\Phi(t),$$

где $t = \frac{\delta\sqrt{n}}{\sigma}$.

Найдя из последнего равенства $\delta = \frac{t\sigma}{\sqrt{n}}$, можно записать

$$P\left(\left|\overline{X} - a\right| < \frac{t\sigma}{\sqrt{n}}\right) = 2\Phi(t).$$

Приняв во внимание, что вероятность P задана и равна γ, окончательно имеем

$$P\left(\overline{x} - \frac{t\sigma}{\sqrt{n}} < a < \overline{x} + \frac{t\sigma}{\sqrt{n}}\right) = 2\Phi(t) = \gamma.$$

Таким образом, с надежностью γ можно утверждать, что доверительный интервал $\left(\overline{x} - \frac{t\sigma}{\sqrt{n}}; \overline{x} + \frac{t\sigma}{\sqrt{n}}\right)$ покрывает неизвестный параметр a; точность оценки $t = \frac{\delta\sqrt{n}}{\sigma}$.

Замечание. Число t определяется из равенства $2\Phi(t) = \gamma$ или $\Phi(t) = \frac{\gamma}{2}$; по таблице функций Лапласа (см. приложение 3) находят аргумент t, которому соответствует значение функции Лапласа, равное $\frac{\gamma}{2}$.

Проведем анализ параметров δ, γ и n.

1. При возрастании объема выборки n число δ убывает и, следовательно, точность оценки увеличивается.

2. Увеличение надежности оценки $\gamma = 2\Phi(t)$ приводит к увеличению t ($\Phi(t)$ возрастающая функция), следовательно, и к возрастанию δ; другими словами, увеличение надежности классической оценки влечет за собой уменьшение ее точности.

3. Если требуется оценить математическое ожидание с наперед заданной точностью δ и надежность γ, то минимальный объем выборки, который обеспечит эту точность, находят по формуле

$$n = \frac{t^2\sigma^2}{\delta^2}$$

(следствие равенства $\delta = \frac{t\sigma}{\sqrt{n}}$).

Доверительный интервал для оценки математического ожидания нормального распределения при неизвестном σ

Пусть количественный признак *X* генеральной совокупности распределен нормально, причем среднее квадратическое отклонение σ неизвестно. Требуется оценить неизвестное математическое ожидание *a* с помощью доверительных интервалов.

Оказывается, что по данным выборки можно построить случайную величину (ее возможные значения будем обозначать через *t*):

$$T = \frac{\overline{X} - a}{s/\sqrt{n}},$$

которая имеет распределение Стьюдента с $k = n - 1$ степенями свободы (см. пояснение в конце параграфа); здесь $\overline{X}$ – выборочная средняя, *s* – «исправленное» среднее квадратическое отклонение, *n* – объем выборки.

Введем новую точность оценки, обозначив ее t_γ (параметр находится по заданным *n* и γ, приложение 4), тогда

$$P\left(\left|\frac{\overline{X} - a}{s/\sqrt{n}}\right| < t_\gamma\right) = \gamma.$$

Заменив неравенство в круглых скобках равносильным ему двойным неравенством, получим

$$P\left(\bar{x} - t_\gamma \frac{s}{\sqrt{n}} < a < \bar{x} + t_\gamma \frac{s}{\sqrt{n}}\right) = \gamma.$$

Итак, пользуясь распределением Стьюдента, мы нашли доверительный интервал $\left(\bar{x} - t_\gamma \frac{s}{\sqrt{n}}; \bar{x} + t_\gamma \frac{s}{\sqrt{n}}\right)$, покрывающий неизвестный параметр *a* с надежностью γ.

Пояснение. Если *Z* – нормальная величина, причем *M(Z)*=0, $\sigma(Z)$=1, а *V* – независимая от *Z* величина, распределенная по закону χ^2 с *k* степенями свободы, то величина

$$T = \frac{Z}{\sqrt{V/k}}$$

распределена по закону Стьюдента с *k* степенями свободы.

Положим случайная величина

$$Z = \frac{\bar{x}_B - a}{\sigma/\sqrt{n}},$$

также имеет нормальное распределение как линейная функция нормального аргумента $\bar{x}_B$ причем *M(Z)*=0, $\sigma(Z)$=1.

Известно, что случайные величины *Z* и $V = \frac{(n-1)S^2}{\sigma^2}$

независимы и что величина V распределена по закону χ^2 с $k=n-1$ степенями свободы. Следовательно, выразив величину T, получим

$$T = \frac{(\overline{x}_B - a)\sqrt{n}}{S},$$

которая распределена по закону Стьюдента с $k=n-1$ степенями свободы.

Доверительные интервалы для оценки среднего квадратического отклонения σ нормального распределения

Пусть количественный признак X генеральной совокупности распределен нормально. Требуется оценить неизвестное генеральное среднее квадратическое отклонение σ по «исправленному» выборочному среднему квадратическому отклонению s. Поставим перед собой задачу найти доверительные интервалы, покрывающие параметр σ с заданной надежностью γ.

Потребуем, чтобы выполнялось соотношение $P(|\sigma - s| < \delta) = \gamma$ или $P(s - \delta < \sigma < s + \delta) = \gamma.$

Для того чтобы можно было пользоваться готовой таблицей, преобразуем двойное неравенство

$$s - \delta < \sigma < s + \delta$$

в равносильное неравенство

$$s\left(1 - \frac{\delta}{s}\right) < \sigma < s\left(1 + \frac{\delta}{s}\right).$$

Положив $\frac{\delta}{s} = q$, получим

$$s(1-q) < \sigma < s(1+q).$$

Параметр q определяется соответствующими n и γ по приложению 5.

Пример. Количественный признак X генеральной совокупности распределен нормально. По выборке объема $n=25$ найдено «исправленное» среднее квадратическое отклонение $s=0,8$. Найти доверительный интервал, покрывающий генеральное среднее квадратическое отклонение σ с надежностью 0,95.

Решение. По таблице приложения 5 по данным $\gamma = 0,95$ и $n = 25$ найдем $q = 0,32$.

Искомый доверительный интервал таков:

$0,8(1-0,32) < \sigma < 0,8(1+0,32)$ или $0,544 < \sigma < 1,056$.

Замечание. Выше предполагалось, что $q<1$. Если $q>1$, то неравенство примет вид (учитывая, что $\sigma>0$)

$$0 < \sigma < s(1+q).$$

Решение типовых задач

Задача 1. Из большой группы предприятий одной из отраслей промышленности случайным образом отобрано 30, по которым получены показатели основных фондов в млн. руб.: 2; 3; 2; 4; 5; 2; 3; 3; 6; 4; 5; 4; 6; 5; 3; 4; 2; 4; 3; 3; 5; 4; 6; 4; 5; 3; 4; 3; 2; 4.

1. Составить дискретное статистическое распределение выборки.

2. Найти объем выборки.
3. Составить распределение относительных частот.
4. Построить полигон частот.
5. Составить эмпирическую функцию распределения и построить ее график.
6. Найти несмещенные оценки числовых характеристик случайной величины.

Решение

1. Расположим различные значения признака в порядке их возрастания и под каждым из них запишем их частоты. Получим дискретное статистическое распределение выборки:

x_i	2	3	4	5	6
n_i	5	8	9	5	3

где x_i - варианты, n_i - частоты вариант x_i.

2. Сумма частот всех вариант должна быть равной объему выборки.

В данном примере объем выборки равен: n=5 + 8 + 9 + 5 + 3=30.

3. Найдем относительные частоты:

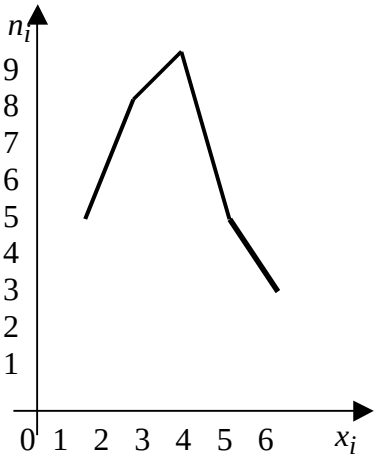
$$W_1=\frac{5}{30}=\frac{1}{6};\ W_2=\frac{8}{30}=\frac{4}{15};\ W_3=\frac{9}{30}=\frac{3}{10};\ W_4=\frac{5}{30}=\frac{1}{6};\ W_5=\frac{3}{30}=\frac{1}{10}.$$

Запишем искомое распределение относительных частот

x_i	2	3	4	5	6
W_i	$\frac{1}{6}$	$\frac{4}{15}$	$\frac{3}{10}$	$\frac{1}{6}$	$\frac{1}{10}$

Контроль: $\frac{1}{6}+\frac{4}{15}+\frac{3}{10}+\frac{1}{6}+\frac{1}{10}=1.$

4. Строим точки с координатами (x_i, n_i) и соединяем их последовательно отрезка-



ми. Полученная ломаная линия называется полигоном частот:

5. Согласно определению эмпирической функцией распределения называется функция вида

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

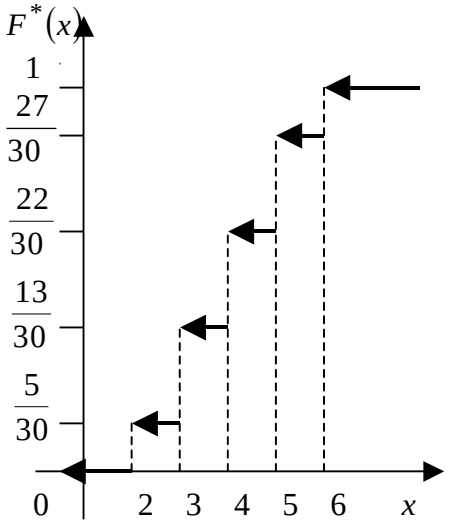
где n – объем выборки; n_x - сумма частот вариант, меньших x.

Эмпирическая функция является оценкой функции распределения генеральной совокупности. Наименьшая варианта равна 2, поэтому при $x \le 2, n_x = 0$ и $F^*(x) = 0$. Значение $X<3$, а именно, $X = x_1 = 2$ наблюдалось 5 раз. Тогда для $2 < x \le 3$ $n_x = 5$ и $F^*(x) = \frac{5}{30}$. Значение $X<4$, а именно, $X=2$, $X=3$, наблюдалось 5 + 8 =13 раз. Поэтому для $3 < x \le 4$ $n_x = 13$ и $F^*(x) = \frac{13}{30}$. Аналогично рассуждая, получаем: для $4 < x \le 5$ $n_x = 5+8+9 = 22$ и $F^*(x) = \frac{22}{30}$, для $5 < x \le 6$ $n_x = 5+8+9+5 = 27$ и $F^*(x) = \frac{27}{30}$ и при $x > 6$ $n_x = 5+8+9+5+3 = 30$ и $F^*(x) = \frac{30}{30} = 1$.

Таким образом,

$$F^*(x) = \begin{cases} 0 \text{ при } & x \le 2; \\ \frac{5}{30} \text{ при } & 2 < x \le 3; \\ \frac{13}{30} \text{ при } & 3 < x \le 4; \\ \frac{22}{30} \text{ при } & 4 < x \le 5; \\ \frac{27}{30} \text{ при } & 5 < x \le 6; \\ 1 \text{ при } & x > 6. \end{cases}$$

График эмпирической функции имеет вид:



6. Несмещенной оценкой математического ожидания является средняя выборочная:

$$\bar{x}_B = \frac{\sum_{i=1}^{k} x_i n_i}{n} = \frac{2 \cdot 5 + 3 \cdot 8 + 4 \cdot 9 + 5 \cdot 5 + 6 \cdot 3}{30} = \frac{113}{30} \approx 3{,}77.$$

Несмещенная оценка дисперсии – исправления выборочная дисперсия:

$$s^2 = \frac{n}{n-1} D_B.$$

$$D_B = \overline{x_B^2} - \left(\bar{x}_B\right)^2 = \frac{2^2 \cdot 5 + 3^2 \cdot 8 + 4^2 \cdot 9 + 5^2 \cdot 5 + 6^2 \cdot 3}{30} - \left(\frac{113}{30}\right)^2 \approx 1{,}42.$$

$$s^2 = \frac{30}{29} \cdot 1{,}42 \approx 1{,}47.$$

Задача 2. Выборочно обследование 30 предприятий машиностроительной промышленности по валовой продукции и получены следующие данные, в млн. руб.:
18,0; 12,0; 11,9; 1,9; 5,5; 14,6; 4,8; 5,6; 4,8; 10,9; 9,7; 7,2; 12,4; 7,6;
9,7; 11,2; 4,2; 4,9; 9,6; 3,2; 8,6; 4,6; 6,7; 8,4; 6,8; 6,9; 17,9; 9,6;
14,8; 15,8.

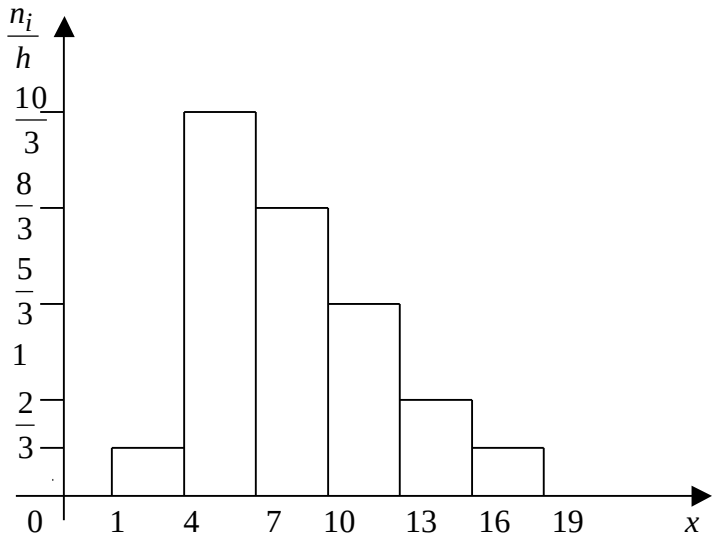
Составить интервальное распределение выборки с началом $x_0 = 1$ и длиной частичного интервала $h = 3$. Построить гистограмму частот.

Решение. Для составления интервального распределения составим таблицу, в первой строке которой расположим в порядке возрастания интервалы, длина каждого из которых $h = 3$. Во второй строке запишем количество значений признака в выборке, попавших в этот интервал (т.е. сумму частот вариант, попавших в соответствующий интервал):

$(x_i; x_{i+1})$	1-4	4-7	7-10	10-13	13-16	16-19
n_i	2	10	8	5	3	2

Объем выборки $n = 2 + 10 + 8 + 5 + 3 + 2 = 30$.

Для построения гистограммы частот на оси абсцисс откладываем частичные интервалы, на каждом из них строим прямоугольники высотой $\frac{n_i}{h}$, где n_i – частота i-го частичного интервала, h – шаг (длина интервала), таким образом, гистограмма примет вид:



Указание. Для построения эмпирической функции распределения и нахождения точечных оценок ряда необходимо преобразовать его к дискретному виду по формуле

$$x_i^* = \frac{x_i + x_{i+1}}{2}.$$

Получим

x_i^*	2,5	5,5	8,5	11,5	14,5	17,5
n_i	2	10	8	5	3	2

Задача 3. Из большой партии электроламп случайным образом отобрано 100. Средняя продолжительность горения ламп в выборке оказалась равной 1000 ч. Найти с надежностью $\gamma = 0{,}95$ доверительный интервал для средней продолжительности a горения ламп во всей партии, если известно, что среднее квадратическое отклонение продолжительности горения лампы $\sigma = 40$ ч и продолжительность горения ламп распределена по нормальному закону.

Решение. По условию $\bar{x}_B = 1000,\ \gamma = 0{,}95,\ \sigma = 40.$ Для решения воспользуемся формулой

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

По приложению 3 находим t из условия:

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

Тогда доверительный интервал:

$$1000 - \frac{1{,}96 \cdot 40}{\sqrt{100}} < a < 1000 + \frac{1{,}96 \cdot 40}{\sqrt{100}}$$

$$992{,}16 < a < 1007{,}84.$$

Задачи (71 – 80)

В задачах 71 – 80 выборочные совокупности заданы из соответствующих генеральных совокупностей. Требуется:

1. Составить интервальное распределения выборки с шагом h, взяв за начало первого интервала x_0.
2. Построить гистограмму частот.
3. Найти $\bar{x}_B; D_B; \sigma_B; S.$
4. Найти с надежностью γ доверительный интервал для оценки неизвестного математического ожидания признака X генеральной совокупности, если признак X распределен по нормальному закону и его среднее квадратическое отклонение равно $\sigma_Г$.

71. Произведено выборочное обследование 25 магазинов по величине товарооборота. Получены следующие результаты (в тыс. руб.):

42,5 60,0 63,5 70,5 82,0 83,5 92,0 95,5 100,0 101,0 105,0 108,5 110,0 115,5 120,0 120,5 122,0 130,0 138,5 140,0 142,0 150,5 160,0 162,1 180,5

$\gamma = 0{,}96;\ \sigma_Г = 31;\ h = 20;\ x_0 = 42{,}5.$

72. Объем промышленного производства Российской Федерации за период с 1999-2000 годы составил (млрд. руб.) в месяц:

187,6 189,8 223,0 223,2 213,2 228,6 242,3 252,7 271,2 293,7 311,8 358,1 331,7 350,8 387,5 359,2 361,1 384,5 391,6 407,7 417,6 442,7 451,9 476,2

$\gamma = 0{,}98;\ \sigma_Г = 86{,}63;\ h = 50;\ x_0 = 185.$

73. Темп роста курса акций 25 фирм по сравнению с предыдущим месяцем составил(%)
104 103,1 102 98 99 94 119 114,8 109,5 103,1 92 97,1 95,2 91,7
104 104,5 92,8 95,8 104,9 77,5 93,1 94,9 99,5 99,7 103
$\gamma = 0{,}95;\ \sigma_Г = 8{,}05;\ h = 10;\ x_0 = 75.$

74. Получены результаты выборочного обследования по выполнению плана выработки на одного рабочего (в %):
90,0 96,0 98,0 98,0 98,5 99,0 101,5 102 102,0 102,5 103,0 103,5 104,0 104,0 104 104,5 105,5 106,0 108,0 108,2 108,7 109,0 112 113,5
$\gamma = 0{,}98;\ \sigma_Г = 4{,}7\%;\ h = 5;\ x_0 = 90.$

75. Были испытаны 25 ламп на продолжительность горения и получены следующие результаты (в часах):
773 792 815 827 843 854 861 869 877 886 889 892 885 901 903 905 911 918 919 923 929 937 941 955 981
$\gamma = 0{,}92;\ \sigma_Г = 50;\ h = 40;\ x_0 = 760.$

76. Для прогнозирования спроса на свою продукцию предприятие проводит исследование, в результате которого получены данные о размере реализованной продукции за некоторый период времени (в тыс. руб.):
42,5 60,0 63,5 70,5 82,0 83,5 92,0 95,5 100,0 101,0 105,0 108,5 110,0 115,5 120,0 130,0 138,5 140,0 142,0 150,5 160,0 162,1 180,5
$\gamma = 0{,}96;\ \sigma_Г = 31;\ h = 30;\ x_0 = 40.$

77. В течение 25 лет наблюдался подъем уровня воды в реке во время паводков. Получены следующие значения (в см.):
266 278 315 336 347 354 368 369 391 408 411 416 427 444 448 457 462 481 483 495 512 536 576
$\gamma = 0{,}96;\ \sigma_Г = 65;\ h = 50;\ x_0 = 20.$

78. На предприятии было произведено выборочное обследование заработной платы рабочих и получены следующие результаты (в руб.)
1360 1550 1600 1690 1750 1750 1800 1880 1890 1920
1950 2000 2020 2050 2050 2050 20801 2120 2150 2200
2250 2340 2420 2450 2600
$\gamma = 0{,}95;\ \sigma_Г = 300\text{руб};\ h = 200;\ x_0 = 1300.$

79. Для определения себестоимости строительно-монтажных работ было произведено выборочное обследование 25 стороительно-монтажных управлений и получены следующие результаты (тыс. руб.)
1250 1450 1550 1700 1760 1820 1880 1960 2100 2175 2190 2200 2220 2275 2280 2310 2400 2550 2580 2600 2670 2800 2950 3000 3075
$\gamma = 0{,}94;\ \sigma_Г = 446\text{руб};\ h = 400;\ x_0 = 1100.$

80. В районной сберегательной кассе проведено выборочное обследование 25 вкладов, которое дало следующие результаты (в руб.):
750 2100 3500 3500 4000 5200 5400 5600 5900 6800 7000 7000 7200 7500 7800 7900 8100 8500 8750 8900 9000 10000 11000 12000 12500
$\gamma = 0{,}95;\ \sigma_Г = 2800\text{руб};\ h = 2000;\ x_0 = 500.$

3.2. Элементы теории корреляции

Функциональная, статистическая и корреляционная зависимости

Зависимость между переменными *X* и *Y* называется ***функциональной***, если существует функция *y=f(x)*, по которой каждому значению $x \in X$ ставится в соответствии единственное значение $y \in Y$.

Однако не всякую зависимость между *X* и *Y* можно представить в виде функции. Иногда одному фиксированному значению *X* соответствует множество значений *Y*.

Статистической называют зависимость, при которой изменение одной из величин влечет изменение распределения другой. В частности, статистическая зависимость проявляется в том, что при изменении одной из величин изменяется среднее значение другой; в этом случае статистическую зависимость называют ***корреляционной***.

Другими словами, корреляционной зависимостью признака *Y* от *X* называют функциональную зависимость ***условного среднего*** $\bar{y}_x$ от *X* , т.е. $\bar{y}_x = f(x)$, где $\bar{y}_x$ – среднее арифметическое наблюдавшихся значений *Y*, соответствующих *X* = *x*. Например, если при $x_1 = 2$ величина *Y* приняла значения $y_1 = 5,\ y_2 = 6,\ y_3 = 10,$ то условное среднее $\bar{y}_{x_1} = \frac{(5+6+1)}{3} = 4.$

Аналогично определяется условное среднее $\bar{x}_y$.

Условным средним $\bar{x}_y$ называется среднее арифметическое наблюдавшихся значений *X*, соответствующих *Y* = *y*.

Уравнение $\bar{y}_x = f(x)$, называется ***выборочным уравнением регрессии*** *Y* на *X*; функцию *f(x)* называют ***выборочной регрессией*** *Y* на *X* , а ее график – ***выборочной линией регрессии*** *Y* на *X*. Аналогично уравнение $\bar{x}_y = \varphi(y)$ называется ***выборочным уравнением регрессии*** *X* на *Y*; функцию *φ(x)* называют ***выборочной регрессией*** *X* на *Y* , а ее график – ***выборочной линией регрессии*** *X* на *Y*.

Основные задачи теории корреляции

1. Установить форму корреляционной связи, т.е. вид функции регрессии (линейная, квадратическая , показательная и пр.).

В случае если обе функции *f(x)* и *φ(x)* линейны, то корреляцию называют линейной, в противном случае нелинейной.

2. Оценить тесноту (силу) корреляционной связи.

Теснота корреляционной связи между *X* и *Y* оценивается по величине рассеяния значений *Y* вокруг условного среднего $\bar{y}_x$. Большое рассеяние свидетельствует о слабой зависимости *Y* от *X* либо об отсутствии этой зависимости. Малое рассеяние указывает на наличие достаточно сильной зависимости; возможно даже, что *X* и *Y* связаны функционально.

Аналогично (по величине рассеяния значений *X* вокруг среднего $\bar{x}_y$) оценивают тесноту корреляционной связи *Y* от *X*.

Корреляционная таблица

Будем считать, что объекты выборочной совокупности характеризуются парой признаков X и Y ,т.е. каждому объекту соответствует пара чисел $(x;y)$. Статистическую совокупность будем обозначать Ω.

Пусть каждому объекту с признаками $(x;y)$ из совокупности Ω поставлена в соответствие точка плоскости XOY с координатами $(x;y)$. Полученное множество точек называется ***диаграммой рассеяния*** статистической совокупности Ω.

Диаграмма рассеяния, как и график функции, дает наглядное представление о статистической совокупности, о виде зависимости между факторами, о тесноте связи.

Если Ω - конечная совокупность объема n, то ее можно описать корреляционной таблицей

Y / X	y_1	$\cdots$	y_j	$\cdots$	y_l	n_x
x_1	n_{11}	$\cdots$	n_{1j}	$\cdots$	n_{1l}	n_{x1}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
x_i	n_{i1}	$\cdots$	n_{ij}	$\cdots$	n_{il}	n_{xi}
$\vdots$	$\vdots$	$\vdots$		$\vdots$		$\vdots$
x_k	n_{k1}	$\cdots$	n_{kj}	$\cdots$	n_{kl}	n_{xk}
n_y	n_{y1}	$\cdots$	n_{yj}	$\cdots$	n_{yl}	n

где $x_i\left(i=\overline{1,k}\right)$, $y_j\left(j=\overline{1,l}\right)$ – соответственно значения признаков X и Y; n_{xi}, n_{yj} - соответствующие им частоты; n_{ij} – частота, с которой встречается пара (x_i, y_j). По определению, $n_{xi}=\sum\limits_{j=1}^{l} n_{ij}$, $n_{yj}=\sum\limits_{i=1}^{k} n_{ij}$. Из таблицы вытекают следующие равенства для объема выборки n: $n=\sum\limits_{i=1}^{k} n_{xi}=\sum\limits_{j=1}^{l} n_{yj}=\sum\limits_{i=1}^{k}\sum\limits_{j=1}^{l} n_{ij}$.

По опытным данным, приведенным в корреляционной таблице, можно судить о форме корреляционной связи между признаками X и Y. С этой целью находятся условные средние $\bar{y}_{xi}$, соответствующие значениям $x_i=\left(i=\overline{1,k}\right)$, и $\bar{x}_{yj}$, соответствующие значениям $y_j=\left(j=\overline{1,l}\right)$ по формулам

$$\bar{y}_{xi}=\frac{\sum\limits_{j=1}^{l} y_j n_{ij}}{n_{xi}};\quad \bar{x}_{yj}=\frac{\sum\limits_{i=1}^{k} x_i n_{ij}}{n_{yi}}.$$

Эмпирической линией регрессии Y на X (X на Y) называют ломаную линию, соединяющую отрезками точки с координатами $M_i^*\left(x_i;\bar{y}_{xi}\right)\left(M_j^*\left(\bar{x}_{yj};y_j\right)\right)$.

Форма полученной таким образом эмпирической ломаной является прообразом формы теоретической зависимости.

Уравнение прямой линии регрессии

Пусть требуется найти теоретическое уравнение $\overline{Y}_x = f(x,\ a_1,\ \ldots, a_n)$ регрессии *Y* на *X*. Параметры $a_1,\ \ldots, a_n$ этого уравнения находятся методом "наименьших квадратов". При предполагаемом законе функциональной зависимости *f*, коэффициенты $a_1,\ \ldots, a_n$ выбирают «наилучшим» образом так, чтобы величина $\left|\overline{Y}_{xi} - \overline{y}_{xi}\right|$ была наименьшей. Данная величина определяет расстояние от точек $M_i\ \left(x_i; \overline{Y}_{xi}\right)$, лежащих на предполагаемой теоретической кривой до угловых точек $M_i^*\left(x_i; \overline{y}_{xi}\right)$ эмпирической кривой.

Допустим, что количественные признаки *X* и *Y* связаны линейной корреляционной зависимостью. В этом случае обе линии регрессии будут прямыми. Тогда $f(x,\ a_1,\ \ldots, a_n) = a_1 x + a_2$, а теоретической кривой *Y* на *X* будет прямая

$$\overline{Y}_x = a_1 x + a_2 \quad . \qquad (*)$$

Найдем параметры a_1 и a_2 так, чтобы точки M_i^*, построенные по данным наблюдений на плоскости *XOY* как можно ближе лежали к теоретической прямой.

Рассмотрим разность $Y_i - y_i\left(i = \overline{(1,n)}\right)$, где Y_i - ордината, соответствующая x_i, вычисляется по равенству (*), y_i - наблюдаемая ордината, соответствующая значению x_i.

Составим функцию, зависящую от параметров a_1 и a_2:

$$F(a_1; a_2) = \sum_{i=1}^{n} (Y_i - y_i)^2 = \sum_{i=1}^{n} \left(a_1 x_i + a_2 + y_i\right)^2.$$

Для отыскания минимума данной функции приравняем к нулю частные производные:

$$F'_{a_1} = 2\sum_{i=1}^{n} (a_1 x_i + a_2 - y_i) x_i = 0;$$

$$F'_{a_2} = 2\sum_{i=1}^{n} (a_1 x_i + a_2 - y_i) = 0.$$

Получим систему уравнений

$$\begin{cases} a_1 \sum\limits_{i=1}^{n} x_i^2 + a_2 \sum\limits_{i=1}^{n} x_i = \sum\limits_{i=1}^{n} x_i y_i; \\ a_1 \sum\limits_{i=1}^{n} x_i + a_2 n = \sum\limits_{i=1}^{n} y_i. \end{cases}$$

Воспользуемся следующими тождествами:

$$\overline{x} = \frac{\sum x_i}{n} \Rightarrow \sum x_i = \overline{x} \cdot n;\ \overline{y} = \frac{\sum y_i}{n} \Rightarrow \sum y_i = \overline{y} \cdot n;$$

$$\overline{x^2}=\frac{\sum x_i^2}{n}\Rightarrow\sum x_i^2=\overline{x^2}\cdot n;\ \ \overline{xy}=\frac{\sum x_i y_i}{n}\Rightarrow\sum x_i y_i=\overline{xy}\cdot n.$$

Подставим полученные выражения в систему и разделим каждое уравнение на n

$$\begin{cases}a_1\overline{x^2}+a_2\bar{x}=\overline{xy};\\ a_1\bar{x}+a_2=\bar{y}.\end{cases}$$

Решением системы являются коэффициенты:

$$a_1=\frac{\overline{xy}-\bar{x}\bar{y}}{\sigma_x^2}\ \text{ и }\ a_2=\bar{y}-a_1\bar{x}.$$

Угловой коэффициент a_1 прямой линии регрессии Y на X называется ***коэффициентом регрессии*** Y на X и обычно обозначается $\rho_{yx}=\frac{\overline{xy}-\bar{x}\bar{y}}{\sigma_x^2}$. Тогда уравнение (∗) можно записать следующим образом:

$$\bar{Y}_x-\bar{y}=\rho_{yx}\left(x-\bar{x}\right).$$

Аналогично теоретическое уравнение $\bar{X}_y=b_1y+b_2$ линейной регрессии X на Y с помощью коэффициента $\rho_{xy}=\frac{\overline{xy}-\bar{x}\bar{y}}{\sigma_y^2}$ приводится к виду

$$\bar{X}_y-\bar{x}=\rho_{xy}\left(y-\bar{y}\right).$$

Сравнивая коэффициенты регрессии Y на X и X на Y, можно отметить, что они имеют одинаковые знаки (в силу совпадения числителей и положительности знаменателей).

Выборочный коэффициент корреляции

Выборочным коэффициентом корреляции r_B признаков X и Y называется число, равное среднему геометрическому коэффициентов регрессии и имеющее их знак:

$$r_B=\pm\sqrt{\rho_{xy}\rho_{yx}}=\frac{\overline{xy}-\bar{x}\bar{y}}{\sigma_x\sigma_y}.$$

Уравнения регрессии с помощью коэффициента корреляции примут вид:

$$\bar{Y}_x-\bar{y}=r_B\frac{\sigma_y}{\sigma_x}\left(x-\bar{x}\right);$$

$$\bar{X}_y-\bar{x}=r_B\frac{\sigma_x}{\sigma_y}\left(y-\bar{y}\right).$$

Прямые регрессии пересекаются в точке $\left(\bar{x};\bar{y}\right)$, которая называется средней точкой корреляционного графика.

Коэффициент корреляции имеет важное самостоятельное значение. С его помощью оценивается теснота (сила) корреляционной связи между признаками. Коэффициент корреляции r_B обладает следующими свойствами:

1. $\left|r_B\right|\le1$ или $-1\le r_B\le1$.

2. Условие $|r_B| = 1 \quad (r_B = \pm 1)$ является необходимым и достаточным условием существования линейной функциональной зависимости.

3. При $r_B = 0$ линейной корреляционной связи между признаками не существует (при этом может быть нелинейная корреляционная связь и даже нелинейная функциональная зависимость).

Таким образом, чем ближе по модулю коэффициент линейной корреляции к единице, чем теснее линейная зависимость между *X* и *Y*, чем ближе коэффициент корреляции к нулю, тем слабее линейная зависимость.

О тесноте связи можно судить по значению коэффициента корреляции, используя шкалу Чеддока:

Показания тесноты связи	0,1-0,3	0,3-0,5	0,5-0,7	0,7-0,9	0,9-0,99
Характеристика силы связи	слабая	умеренная	заметная	высокая	весьма высокая

Если выборка имеет достаточно большой объем $n \geq 50$ и является репрезентативной, то заключение о тесноте связи признаков *X* и *Y* может быть распространено на всю генеральную совокупность.

Так, для оценки коэффициента корреляции $r_Г$ нормально распределенной совокупности можно использовать формулу

$$r_B - 3\frac{1-r_B^2}{\sqrt{n}} \leq r_Г \leq r_B + 3\frac{1+r_B^2}{\sqrt{n}}.$$

Уравнения регрессии в случае равноотстоящих значений признаков

В случае, если значения хотя бы одного из признаков являются равностоящими, полезно использовать условные варианты.

Пусть для определенности значения признака *X* являются равноотстоящими. Тогда расчет основных параметров уравнения регрессии производится по алгоритму:

1. Рассчитываем условные варианты

$$u_i = \frac{x_i - C}{h},$$

где *C* – ложный нуль, *h* – шаг.

2. Находим условные эмпирические моменты первого и второго порядка:

$$M_1^* = \frac{\sum u_i n_i}{n} = \overline{u}; \quad M_2^* = \frac{\sum u_i^2 n_i}{n} = \overline{u^2}.$$

3. Находим

$$\overline{x} = M_1^* h + C = \overline{u}h + C;$$

$$\sigma_x = \sqrt{\overline{u^2} - (\overline{u})^2} \cdot h = \sigma_u \cdot h.$$

4. Вычисляем $r_B = \dfrac{\overline{uy} - \bar{u} \cdot \bar{y}}{\sigma_u \cdot \sigma_y}$.

Пусть значения обоих признаков X и Y являются равноотстоящими соответственно с шагом h_1 и h_2. Тогда целесообразно воспользоваться следующим алгоритмом:

1. Переходим к условным вариантам

$$u_i = \frac{x_i - C_1}{h_1};\ v_i = \frac{y_j - C_2}{h_2}.$$

2. Находим условные эмпирические моменты

$$\bar{u} \text{ и } \bar{v};\ \overline{u^2} \text{ и } \overline{v^2}.$$

3. Находим

$$\bar{x} = \bar{u}h_1 + C_1;\ \sigma_x = \sigma_u \cdot h_1;$$
$$\bar{y} = \bar{v}h_2 + C_2;\ \sigma_y = \sigma_v \cdot h_2$$

4. Коэффициент корреляции определяется по формуле

$$r_B = \frac{\overline{uv} - \bar{u} \cdot \bar{v}}{\sigma_u \cdot \sigma_v}.$$

5. Составляем уравнения регрессии.

Криволинейная корреляция

Между признаками X и Y могут существовать и нелинейные корреляционные зависимости (параболическая, гиперболическая, показательная и пр.).

Рассмотрим подробнее случаи параболической и гиперболической зависимости. Предположим между признаками X и Y – параболическая корреляционная связь. Тогда уравнения регрессии имеют вид:

$$\bar{Y}_x = a_1 x^2 + a_2 x + a_3;$$
$$\overline{X}_y = b_1 y^2 + b_2 y + b_3.$$

Основываясь на выше описанном методе «наименьших квадратов», получим следующую систему линейных уравнений для нахождения параметров:

$$\begin{cases} a_1\overline{x^4} + a_2\overline{x^3} + a_3\overline{x^2} = \overline{x^2 y}; \\ a_1\overline{x^3} + a_2\overline{x^2} + a_3\bar{x} = \overline{xy}; \\ a_1\overline{x^2} + a_2\bar{x} + a_3 = \bar{y}. \end{cases} \qquad (*)$$

Решением системы (*) являются «наилучшие» параметры искомой параболы. Для нахождения параметров b_1, b_2, b_3 необходимо составить идентичную систему уравнений.

В случае гиперболической корреляционной зависимости Y от X уравнения регрессии имеют вид:

$$\bar{Y}_x = \frac{a_1}{x} + a_2;\ \overline{X}_y = \frac{b_1}{y} + b_2.$$

Метод "наименьших квадратов" приводит процесс составления уравнения регрессии к решению следующей системы:

$$\begin{cases} a_1\left(\overline{\frac{1}{x^2}}\right)+a_2\left(\overline{\frac{1}{x}}\right)=\left(\overline{\frac{1}{x}y}\right); \\ a_1\left(\overline{\frac{1}{x}}\right)+a_2=\overline{y}. \end{cases}$$

Аналогично составляется и решается система уравнений относительно параметров b_1 и b_2.

Для оценки тесноты нелинейной корреляционной связи используют ***выборочные корреляционные отношения***: η_{yx} – выборочное корреляционное отношение *Y* к *X*; η_{xy} – выборочное корреляционное отношение *X* к *Y*.

Выборочным корреляционным отношением *Y* к *X* называют отношение межгруппового среднего квадратического отклонения к общему среднему квадратическому отклонению признака *Y*:

$$\eta_{yx}=\frac{\sigma_{\text{межгр}}}{\sigma_{\text{общ}}}$$

или в других обозначениях

$$\eta_{yx}=\frac{\sigma_{\overline{y}_x}}{\sigma_y}.$$

Здесь

$$\sigma_{\overline{y}_x}=\sqrt{\frac{\sum n_x\left(\overline{y}_x-\overline{y}\right)^2}{n}};\ \sigma_y=\sqrt{\frac{\sum n_y\left(y-\overline{y}\right)^2}{n}},$$

где n – объем выборки (сумма всех частот); n_x - частота значения x признака *X*; n_y - частота значения y признака *Y*; $\overline{y}$ - общая средняя признака *X*; $\overline{y}_x$ - условная средняя признака *Y*.

Аналогично определяется выборочное корреляционное отношение *X* к *Y*:

$$\eta_{xy}=\frac{\sigma_{\overline{x}_y}}{\sigma_x}.$$

Свойства выборочного корреляционного отношения

Поскольку η_{xy} обладает тем же свойством, что и η_{yx}, перечислим свойства только выборочного корреляционного отношения η_{yx}, которое далее для упрощения записи будем обозначать через η и для простоты называть «корреляционным отношением».

С в о й с т в о 1. Корреляционное отношение удовлетворяет двойному неравенству

$$0\le\eta\le 1.$$

С в о й с т в о 2. Если $\eta=0,$ то признак *Y* с признаком *X* корреляционной зависимостью не связан.

С в о й с т в о 3. Если $\eta = 1$, то признак Y связан с признаком X функциональной зависимостью.

С в о й с т в о 4. Выборочное корреляционное отношение не меньше абсолютной величины выборочного коэффициента корреляции: $\eta \geq |r_B|$.

С в о й с т в о 5. Если выборочное корреляционное отношение равно абсолютной величине выборочного коэффициента корреляции, то имеет место точная линейная корреляционная зависимость.

Другими словами, если $\eta = |r_B|$, то точки $(x_1; y_1)$, $(x_2; y_2)$, …, $(x_n; y_n)$ лежат на прямой линии регрессии, найденной способом "наименьших квадратов".

Понятие множественной корреляции

Множественная корреляция – это исследование связи между несколькими признаками.

Пусть Z линейно зависит от X и Y, тогда уравнение линейной множественной регрессии имеет вид:

$$z = a_1 x + a_2 y + a_0. \qquad (*)$$

Коэффициенты множественной регрессии a_1, a_2, и a_0 находятся методом "наименьших квадратов", т.е. так, чтобы функция $F(a_1, a_2, a_0) = \sum_i (a_1 x_i + a_2 y_i + a_0 - z_i)^2 n_i$ имела минимум.

Раскрывая знак суммы и группируя слагаемые, приводим уравнение (*) к виду:

$$z - \bar{z} = a_1(x - \bar{x}) + a_2(y - \bar{y}),$$

причем коэффициенты регрессии определяются равенствами:

$$a_1 = \frac{r_{xz} - r_{yz} \cdot r_{xy}}{1 - r_{xy}^2} \cdot \frac{\sigma_z}{\sigma_x};\ a_2 = \frac{r_{yz} - r_{xz} \cdot r_{xy}}{1 - r_{xy}^2} \cdot \frac{\sigma_z}{\sigma_y},$$

где r_{xz}; r_{yz} и r_{xy} – коэффициенты корреляции соответственно между признаками X и Z; Y и Z; X и Y.

Теснота линейной корреляционной связи признака Z с X и Y оценивается с помощью ***выборочного совокупного коэффициента корреляции:***

$$R = \sqrt{\frac{r_{xz}^2 - 2 r_{xy} \cdot r_{xz} \cdot r_{yz} + r_{yz}^2}{1 - r_{xy}^2}}.$$

При этом $0 \leq R \leq 1$ и при приближении R к единице теснота линейной связи Z с X и Y увеличивается.

Следующей задачей множественной корреляции является задача оценить влияние на Z отдельно признака X и отдельно признака Y. Это осуществляется при помощи ***выборочных частных коэффициентов корреляции:***

$$r_{xz(y)} = \frac{r_{xz} - r_{xy} \cdot z_{yz}}{\sqrt{\left(1 - r_{xy}^2\right)\left(1 - r_{yz}^2\right)}};\ r_{yz(x)} = \frac{r_{yz} - r_{xy} \cdot r_{xz}}{\sqrt{\left(1 - r_{xy}^2\right)\left(1 - r_{xz}^2\right)}}.$$

Первый коэффициент оценивает тесноту линейной корреляционной связи между Z и X, когда Y остается постоянным. Теснота связи между Z и Y (при постоянной X) оценивается вторым коэффициентом корреляции $r_{yz(x)}$.

Эти коэффициенты имеют те же свойства, что и обыкновенный выборочный коэффициент корреляции.

Решение типовых задач

Задача 1. Выборочно обследовано 100 заводов по величине основных производственных фондов X (млн. руб.) и объему готовой продукции Y (млн. руб.). Результаты представлены в корреляционной таблице (табл. 1).

Таблица 1

Y	X					n_y
	5	15	25	35	45	
30	7	1				8
32	2	7	1			10
34	1	5	4	1		11
36		1	15	10	8	34
38			3	12	15	30
40				1	6	7
n_x	10	14	23	24	29	n=100

По данным исследования требуется:

1) в прямоугольной системе координат построить эмпирические ломаные регрессии Y на X и X на Y;

2) оценить тесноту линейной корреляционной связи;

3) составить линейные уравнения регрессии Y на X и X на Y и построить их графики в одной системе координат.

Решение. 1. Так как при x = 5 признак Y имеет распределение

Y	30	32	34
n_i	7	2	1

то условное среднее $\overline{y}_{x=5} = \dfrac{30\cdot 7 + 32\cdot 2 + 34\cdot 1}{10} = 30{,}8.$

При x=15 признак Y имеет распределение

Y	30	32	34	36
n_i	1	7	5	1

Следовательно $\overline{y}_{x=15} = \dfrac{30\cdot 1 + 32\cdot 7 + 34\cdot 5 + 36\cdot 1}{14} = 32{,}86.$

Аналогично вычисляются все условные средние $\overline{y}_x$. В результате получим таблицу, выражающую корреляционную зависимость $\overline{y}$ от X (табл. 2).

Таблица 2

X	5	15	25	35	45
$\bar{y}_x$	30,8	32,86	35,74	37,08	37,86

Так как при y=30 признак X имеет распределение

X	5	15
n_j	7	1

то условное среднее $\bar{x}_{y=30} = \frac{5 \cdot 7 + 15 \cdot 1}{8} = 6{,}25.$

При y = 32 признак X имеет распределение

X	5	15	25
n_j	2	7	1

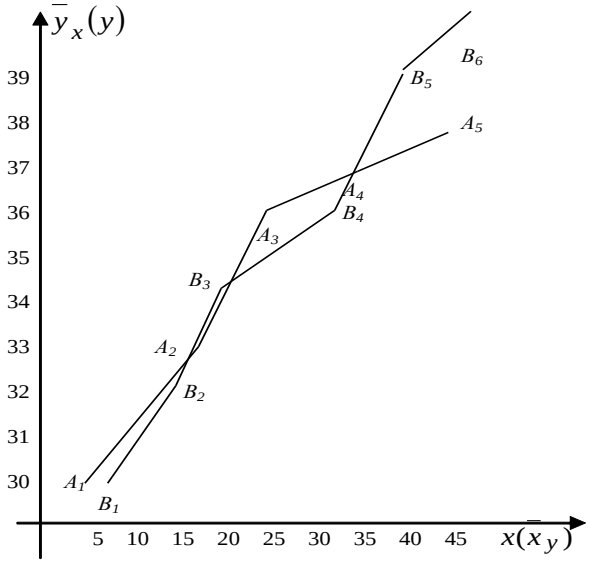
Следовательно $\bar{x}_{y=32} = \frac{5 \cdot 2 + 15 \cdot 7 + 25 \cdot 1}{10} = 14.$

Аналогично вычисляются все $\bar{x}_y$. В результате получим табл. 3.

Таблица 3

Y	30	32	34	36	38	40
$\bar{x}_y$	6,25	14	19,54	32,35	39	43,57

В прямоугольной системе координат построим точки $A_i\left(x_i; \bar{y}_{x_i}\right)$, соединим их отрезками прямых, получим эмпирическую линию регрессии Y на X. Аналогично строятся точки



$B_j\left(\bar{x}_y; y_j\right)$ и эмпирическая линия регрессии X на Y.

2. Выдвинув гипотезу о линейной корреляционной зависимости, оценим тесноту связи. Вычислим выборочный коэффициент корреляции

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

$$\bar{x} = \frac{\sum x_i \cdot n_i}{n},\ \bar{y} = \frac{\sum y_j \cdot n_j}{n},\ \overline{x^2} = \frac{\sum x_i^2 \cdot n_i}{n},\ \overline{y^2} = \frac{\sum y_j^2 \cdot n_j}{n},$$

$$\overline{xy} = \frac{\sum x_i y_j \cdot n_{ij}}{n},\ \sigma_x = \sqrt{\overline{x^2} - (\bar{x})^2}\ ,\ \sigma_y = \sqrt{\overline{y^2} - (\bar{y})^2}$$

$$\bar{x} = \frac{5 \cdot 10 + 15 \cdot 14 + 25 \cdot 23 + 35 \cdot 24 + 45 \cdot 29}{100} = 29{,}8;$$

$$\bar{y} = \frac{30 \cdot 8 + 32 \cdot 10 + 34 \cdot 11 + 36 \cdot 34 + 38 \cdot 30 + 40 \cdot 7}{100} = 35{,}78;$$

$$\overline{x^2} = \frac{5^2 \cdot 10 + 15^2 \cdot 14 + 25^2 \cdot 23 + 35^2 \cdot 24 + 45^2 \cdot 29}{100} = 1059;$$

$$\overline{y^2} = \frac{30^2 \cdot 8 + 32^2 \cdot 10 + 34^2 \cdot 11 + 36^2 \cdot 34 + 45^2 \cdot 30 + 40^2 \cdot 7}{100} = 1287{,}4$$

$$\overline{xy} = \frac{30 \cdot 5 \cdot 7 + 30 \cdot 15 \cdot 1 + 32 \cdot 5 \cdot 2 + 32 \cdot 15 \cdot 7 + 32 \cdot 25 \cdot 1 + 34 \cdot 5 \cdot 1 + 34 \cdot 15 \cdot 5}{100} +$$

$$+ \frac{34 \cdot 25 \cdot 4 + 34 \cdot 35 \cdot 1 + 36 \cdot 15 \cdot 1 + 36 \cdot 25 \cdot 15 + 36 \cdot 35 \cdot 10 + 36 \cdot 45 \cdot 8 + 38 \cdot 25 \cdot 3}{100} +$$

$$+ \frac{38 \cdot 35 \cdot 12 + 38 \cdot 45 \cdot 15 + 40 \cdot 35 + 40 \cdot 45 \cdot 6}{100} = 1095{,}5$$

$$\sigma_x = \sqrt{1059 - (29{,}8)^2} = 13{,}08;\ \sigma_y = \sqrt{1287{,}4 - (35{,}78)^2} = 2{,}68;$$

$$r_B = \frac{1095{,}5 - 29{,}8 \cdot 35{,}78}{13{,}08 \cdot 2{,}68} = 0{,}83.$$

Так как r_B близок к единице, то между Y и X имеется достаточно тесная корреляционная связь.

3. Подставляя найденные величины в уравнения

$$\bar{Y}_x - \bar{y} = r_B \frac{\sigma_y}{\sigma_x}(x - \bar{x}),\ \bar{X}_y - \bar{x} = r_B \frac{\sigma_x}{\sigma_y}(y - \bar{y}),$$

получаем искомые уравнения регрессии:

1) уравнение регрессии Y на X

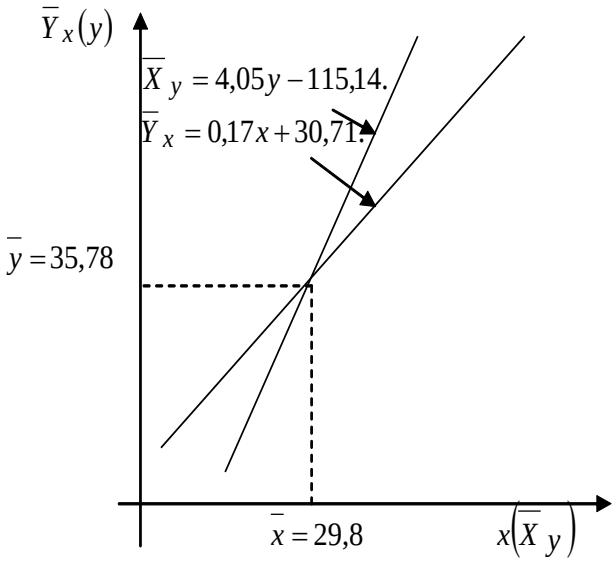
$$\bar{Y}_x - 35{,}78 = 0{,}83 \frac{2{,}68}{13{,}08}(x - 29{,}8),\ \bar{Y}_x = 0{,}17x + 30{,}71.$$

2) уравнение регрессии X на Y

$$\bar{X}_y - 29{,}8 = 0{,}83 \frac{13{,}08}{2{,}68}(y - 35{,}78),\ \bar{X}_y = 4{,}05y - 115{,}14.$$

Замечание. Если в корреляционной таблице даны интервальные распределения, то за значения вариант нужно брать середины частичных интервалов.

Изобразим графики прямых линий регрессии на чертеже.



Так как значения признаков X и Y являются равноотстоящими, то можно данную задачу решить с помощью условных вариант.

Так, в данном примере

$$C_1 = 25,\ h_1 = 10,\ u_i = \frac{x_i - 25}{10};$$

$$C_2 = 36,\ h_2 = 2,\ \nu_j = \frac{y_j - 36}{2}.$$

ν	u					n_y
	-2	-1	0	1	2	
-3	7	1				8
-2	2	7	1			10
-1	1	5	4	1		11
0		1	15	10	8	34
1			3	12	15	30
2				1	6	7
n_x	10	14	23	24	29	n=100

$$\overline{u} = \frac{-2\cdot 10 \cdot 14 + 0\cdot 23 + 1\cdot 24 + 2\cdot 29}{100} = 0{,}48;$$

$$\overline{\nu} = \frac{-3\cdot 8 - 2\cdot 10 - 1\cdot 11 + 1\cdot 30 + 2\cdot 7}{100} = -0{,}11;$$

$$\overline{u^2} = \frac{4\cdot 10 + 1\cdot 14 + 1\cdot 24 + 4\cdot 29}{100} = 1{,}94;\ \overline{\nu^2} = \frac{9\cdot 8 + 4\cdot 10 + 1\cdot 11 + 1\cdot 30 + 4\cdot 7}{100} = 1{,}81;$$

$$\overline{u\nu} = \frac{(-3)(-2)\cdot 7 + (-3)(-1)\cdot 1 + (-2)(-2)\cdot 2 + (-2)(-1)\cdot 7 + (-1)(-2)\cdot 1 + (-1)(-1)\cdot 5 + (-1)\cdot 1\cdot 1}{100} +$$

$$+ \frac{1\cdot 1\cdot 12 + 1\cdot 2\cdot 15 + 2\cdot 1\cdot 1 + 2\cdot 2\cdot 6}{100} = 1{,}4;$$

$$\sigma_u = \sqrt{\overline{u^2} - \left(\overline{u}\right)^2} = \sqrt{1{,}94 - 0{,}2304} = 1{,}308,\ \ \sigma_v = \sqrt{1{,}81 - 0{,}012} = 1{,}34;$$

$$r_B = \frac{1{,}4 - 0{,}48 \cdot \left(-0{,}11\right)}{1{,}31 \cdot 1{,}34} = 0{,}83;$$

$$\overline{x} = \overline{u}h_1 + C_1 = 0{,}48 \cdot 10 + 25 = 29{,}8,\ \ \overline{y} = \overline{v}h_2 + C_2 = -0{,}11 \cdot 2 + 36 = 35{,}78;$$

$$\sigma_x = \sigma_u \cdot h_1 = 1{,}308 \cdot 10 = 13{,}08,\ \ \sigma_y = \sigma_v \cdot h_2 = 1{,}34 \cdot 2 = 2{,}68.$$

Подставляя полученные данные в уравнение регрессии, получим

$$\overline{Y}_x = 0{,}17x + 30{,}71;\ \ \overline{X}_y = 4{,}05y - 115{,}14.$$

Задачи (81 – 90)

В задачах 81 – 90 по корреляционной таблице требуется:

1. В прямоугольной системе координат построить эмпирические ломаные регрессии *Y* на *X* и *X* на *Y*, сделать предположение о виде корреляционной связи.
2. Оценить тесноту линейной корреляционной связи.
3. Составить линейные уравнения регрессии *Y* на *X* и *X* на *Y*, построить их графики.

81. В таблице дано распределение объема производственных фондов *X* (млн руб.) и объема выпуска готовой продукции однотипных предприятий *Y* (млн руб.).

Y	*X*						n_y
	12	17	22	27	32	37	
25	2	4					6
35		6	3				9
45			6	35	4		45
55			2	8	6		16
65				14	7	3	24
n_x	2	10	11	57	17	3	*n*=100

82. В таблице дано распределение 55 компаний по возрасту сотрудников *X* и заработной плате *Y* (усл. ден. ед.).

Y	X					n_y
	20-30	30-40	40-50	50-60	70-80	
50-80	5	4				9
80-110		12	8	1		21
110-140			5	5		10
140-170			4	7		11
170-200				2	1	3
200-230					1	1
n_x	5	16	17	15	2	*n*=55

83. В таблице распределение 50 предприятий оптовой торговли по размерам торговой площади *X* (кв. км.) и объемам реализации *Y* (млн руб.).

Y	*X*					n_y
	1-1,5	1,5-2	2-2,5	2,5-3	3-3,5	

5-10	2	1				3
10-15	3	4	3	1		11
15-20		5	10	8		23
20-25			1	6	1	8
25-30				1	4	5
n_x	5	10	14	16	5	n=50

84. В таблице дано распределение 100 однотипных предприятий по основным фондам X (млн руб.) и себестоимости единицы продукции Y (руб.).

Y	X					n_y
	20	30	40	50	60	
1	8	2				10
3	12	20	8			40
5			10	1		11
7			9	6	2	17
9			10	4	8	22
n_x	20	22	37	11	10	n=100

85. В таблице дано распределение 65 заводов по производству продукции X (тыс. ед.) и уровню механизации труда Y (%).

Y	X					n_y
	320-370	370-420	420-470	470-520	520-570	
5-20	2	3				5
20-35	1	6	7	1		15
35-50		3	10	9	2	24
50-65			5	4	6	15
65-80			2	3	1	6
n_x	3	12	24	17	9	n=65

86. В таблице дано распределение 50 предприятий по объему выпуска продукции X (млн руб.) и численности занятых на предприятии Y (чел.).

Y	X					n_y
	2-4	4-6	6-8	8-10	10-12	
30-70	3	4				7
70-110		9	8	1		18
110-150			5	4	1	10
150-190			4	7	2	13
190-230				1	1	2
n_x	3	13	17	13	4	n=50

87. По совокупности 100 предприятий торговли изучается зависимость между ценой товара X (тыс. руб.) и прибылью торгового предприятия Y (млн руб.).

Y	X						n_y
	5	10	15	20	25	30	
45	2	4					6
55		3	5				8
65			5	35	5		45
75			2	8	17		27
85				4	7	3	14
n_x	2	7	12	47	29	3	n=100

88. В таблице дано распределение 100 предприятий, производящих однородную продукцию, по объему производства X (млн руб.) и себестоимости единицы продукции Y (тыс. руб.).

Y	X					n_y
	0,4-1,4	1,4-2,4	2,4-3,4	3,4-4,4	4,4-5,4	
4-6				2	6	8
6-8			4	7	4	15
8-10	1	1	7	5		14
10-12	2	4	1			7
12-14	3	3				6
n_x	6	8	12	4	10	n=100

89. В таблице дано распределение 50 предприятий по потреблению материалов X (т.) и объему произведенной продукции Y (тыс. ед.).

Y	X					n_y
	9	11	13	15	17	
8	2	6				8
9		4	7	4		15
10		5	7	1	1	14
11			2	4	1	7
12				3	3	6
n_x	2	15	16	12	5	n=50

90. В таблице дано распределение 60 предприятий по стоимости основных производственных фондов X (млн руб.) и объему выпуска продукции Y (млн руб.).

Y	X					n_y
	0-2	2-4	4-6	6-8	8-10	
0-0,2	2	2				4
0,2-0,4	2	7	10			19
0,4-0,6		2	17	7		26
0,6-0,8			4	3	2	9
0,8-1,0					2	2

n_x	4	11	31	10	4	n=60

3.3. Проверка статистических гипотез

Статистическая гипотеза. Статистический критерий

Любую задачу, связанную с анализом статистических данных на языке принятия решений можно представить в виде следующего алгоритма:

1. Сбор статистического материала (выборка).
2. Анализ полученных данных.
3. Выдвижение статистической гипотезы.
4. Проверка выдвинутой гипотезы.
5. Принятие решения.

Такая схема действий является универсальной не только в области статистических исследований.

Решения, принятые на основе анализа статистических данных, называются ***статистическими решениями***. Очевидно, они носят вероятностный характер, поскольку сама выборка является случайной, поэтому принятие статистических решений связано с определенным риском.

Рассмотрим подробнее каждый этап в предложенном алгоритме.

1. Сбор данных

Организация выборки и проведение ее исследования подробно разобраны в курсе общей теории статистики. В целом этот этап зависит от характера произведенной выборки (серийная, повторная и т.д.), ее объема, системы единиц для измеряемого признака.

2. Анализ полученных данных

Первичная статистическая информация представляет собой набор значений признака, т.е. некое числовое множество. Основная задача второго этапа – представить это множество значений в форме, приемлемой для дальнейшего выдвижения гипотезы. Для этой цели служат группировка данных в вариационные ряды и построение полигона и гистограммы относительных частот.

Приведем примеры некоторых, наиболее распространенных видов гистограмм (рис 3.3).

Из закона больших чисел в форме Бернулли известно, что при увеличении объема выборки и одновременном измельчении интервалов контур гистограммы приближается к функции плотности. На этом факте и основывается следующий этап - выдвижение гипотезы.

4. Выдвижение гипотезы

Часто по эмпирическому распределению выборки можно выдвинуть предположение о теоретическом распределении всей генеральной совокупности. Если же закон распределения известен, а его параметры нет, то можно предположить их величину.

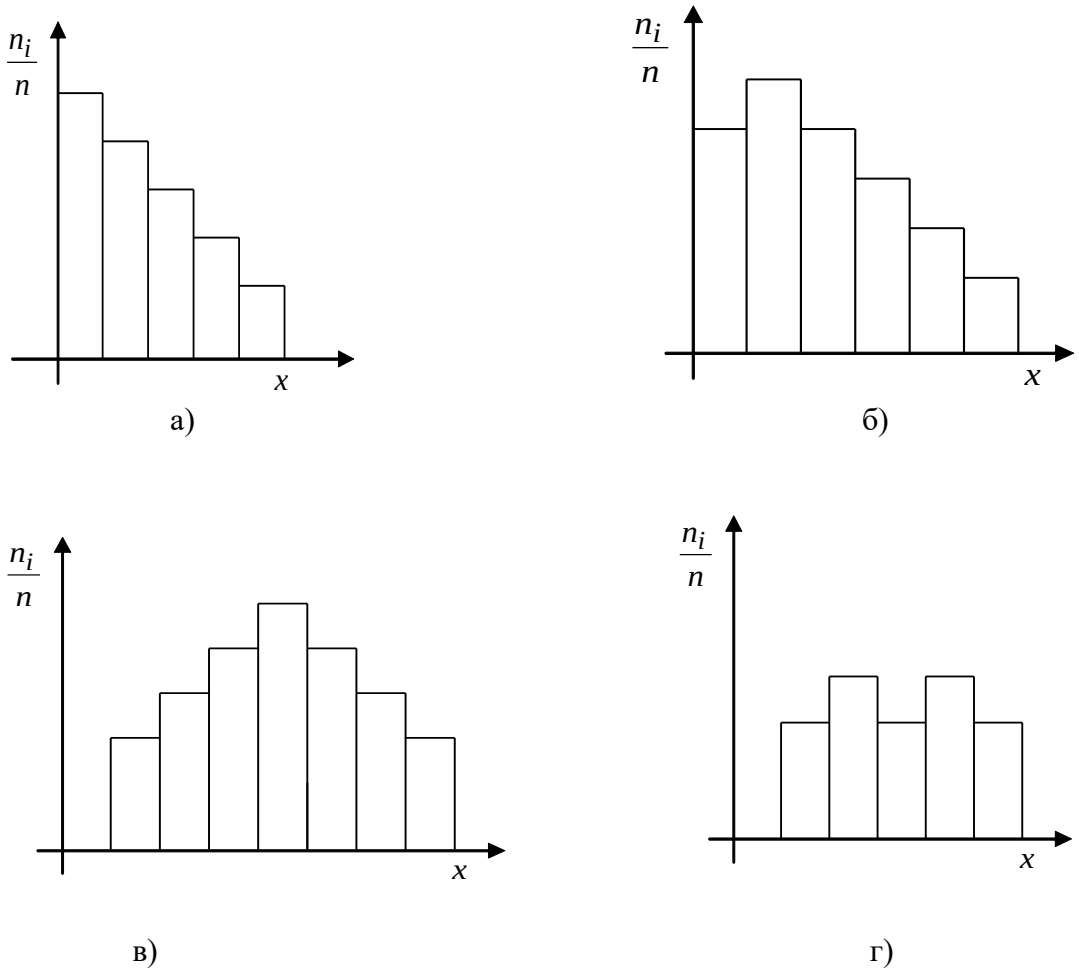


Рис. 3.3.

Статистической называют гипотезу о виде неизвестного распределения, или о параметрах известных распределений.

Например, статистическими являются гипотезы:

1) генеральная совокупность распределена по закону Пуассона;

2) дисперсии двух нормальных совокупностей равны между собой.

В первой гипотезе сделано предположение о виде неизвестного распределения, во второй – о параметрах двух известных распределений.

Наряду с выдвинутой гипотезой рассматривают и противоречащую ей гипотезу. Если выдвинутая гипотеза будет отвергнута, то имеет место противоречащая гипотеза. По этой причине эти гипотезы целесообразно различать.

Нулевой (основной) называют выдвинутую гипотезу H_0.

Конкурирующей (альтернативной) называют гипотезу H_1, которая противоречит нулевой.

Например, если нулевая гипотеза состоит в предположении, что математическое ожидание a нормального распределения равно 10, то конкурирующая гипотеза, в частности, может состоять в предположении, что $a \neq 10$. Коротко это записывают так: H_0 : a=10; H_1: $a \neq 10$.

Различают гипотезы, которые содержат только одно и более одного предположений.

Простой называют гипотезу, содержащую только одно предположение.

Сложной называют гипотезу, которая состоит из конечного или бесконечного числа простых гипотез.

Для того чтобы выдвинуть гипотезу о том или ином виде теоретического распределения, напомним некоторые из них, изученные в курсе теории вероятностей.

1. Распределение Пуассона (рис. 3.4).

$$P_n(m)=\frac{\lambda^m}{m!}e^{-\lambda},\ m=\overline{(0,n)};\quad M(X)=D(X)=\lambda.$$

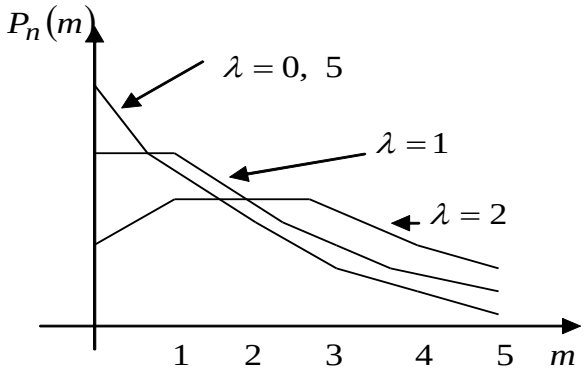


Рис. 3.4.

1. Равномерное распределение (рис. 3.5)

$$f(x)=\begin{cases}0, & x\notin(a;b)\\ \dfrac{1}{b-a}, & x\in[a;b]\end{cases};\quad M(X)=\frac{a+b}{2};\ D(X)=\frac{(b-a)^2}{12}.$$

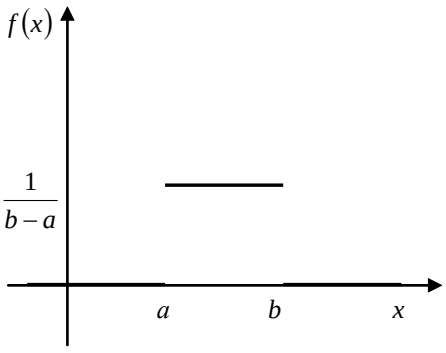


Рис. 3.5.

3. Показательное распределение (рис. 3.6)

$$f(x)=\begin{cases} 0, & x<0 \\ \lambda e^{-\lambda x}, & x\geq 0 \end{cases};\ M(X)\frac{1}{\lambda};D(X)=\frac{1}{\lambda^2}.$$

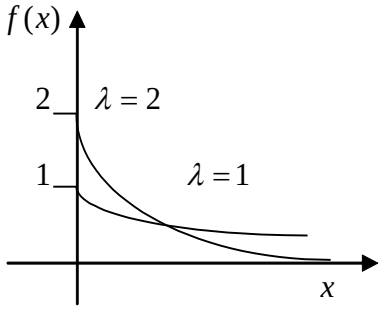


Рис. 3.6.

4. Нормальное распределение (рис. 3.7)

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-a)^2}{2\sigma^2}};\ M(X)=a;D(X)=\sigma^2.$$

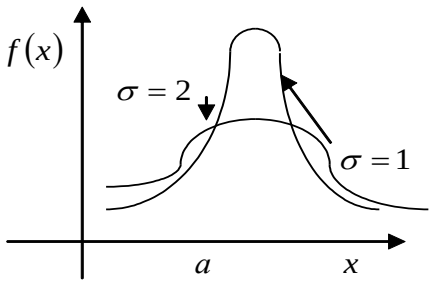


Рис. 3.7.

Сравнивая вид гистограмм приведенных на рис. 3.3, с графиками основных теоретических распределений можно выдвинуть гипотезу о виде распределения всей генеральной совокупности.

Естественно предположить, что на рис. 3.3 (а) – показательное распределение, на рис. 3.3 (б) – распределение Пуассона, на рис.3.3 (в) – нормальное и наконец на рис. 3.3 (г) – равномерное. Итак, по полученной гистограмме выбирается подходящие теоретическое

распределение. Числовые характеристики оцениваются соответствующими выборочными характеристиками.

5. Проверка гипотезы

Проверка гипотезы состоит в том, чтобы установить: можно ли считать расхождение между предполагаемым теоретическим и эмпирическим распределениями несущественным или же существуют коренные (принципиальные) различия? Ответ на этот вопрос дает статистический критерий.

Статистическим критерием называют случайную величину K, которая служит для проверки нулевой гипотезы.

Пусть выдвинута гипотеза о том, что генеральная совокупность распределена нормально и функция плотности имеет вид (рис. 3.8).

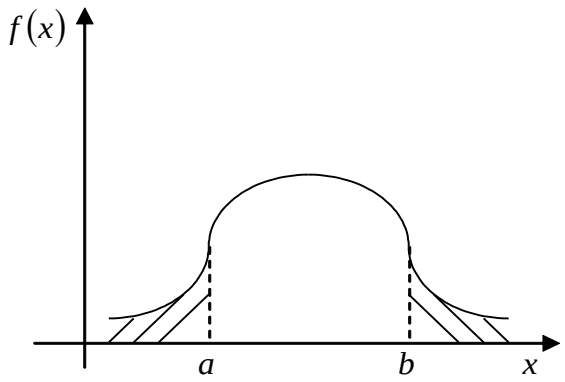


Рис. 3.8.

Тогда можно с уверенностью утверждать, что попадание выборки в заштрихованную область маловероятно, и, напротив, попадание в интервал $(a;b)$ имеет большую вероятность.

Таким образом, если ***наблюдаемое значение критерия*** $K_{набл.}$ (вычисленное по выборке) попадает в интервал $(a;b)$, то это не противоречит гипотезе, если же оно попадает в заштрихованную область, то это ставит гипотезу под сомнение. Следовательно, множество всех возможных значений критерия можно разделить на два непересекающихся множества: одно из них содержит значения критерия, при которых нулевая гипотеза отвергается, а другая – при которых она принимается.

Критической областью называют совокупность значений критерия, при которых нулевую гипотезу отвергают.

Областью принятия гипотезы (областью допустимых значений) называют совокупность значений критерия, при которых нулевую гипотезу принимают.

Основной принцип проверки статистических гипотез можно сформулировать так: если наблюдаемое значение критерия принадлежит критической области – гипотезу отвергают, если наблюдаемое значение критерия принадлежит области принятия гипотезы – гипотезу принимают.

Поскольку критерий K – одномерная случайная величина, все ее возможные значения принадлежат некоторому интервалу. Поэтому критическая область и область принятия

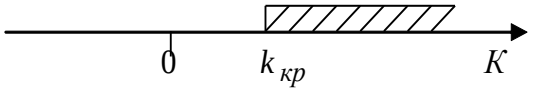
гипотезы также являются интервалами и, следовательно, существуют точки, которые их разделяют.

Критическими точками (границами) $k_{кр}$ называют точки, отделяющие критическую область от области принятия гипотезы.

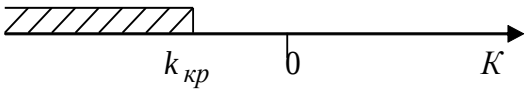
Различают одностороннюю (правостороннюю или левостороннюю) и двустороннюю критические области.

Правосторонней называют критическую область, определяемую неравенством

$$K > k_{кр}, \text{ где } k_{кр} > 0.$$



Левосторонней называют критическую область, определяемую неравенством

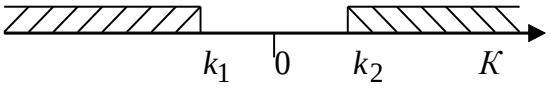


$$K > k_{кр}, \text{ где } k_{кр} < 0.$$

Односторонней называют правостороннюю или левостороннюю критическую область.

Двусторонней называют критическую область, определяемую неравенствами

$$K < k_1, K > k_2, \text{ где } k_2 > k_1.$$



При проверке выдвинутой гипотезы можно допустить два вида ошибок.

1. Если $K_{набл.}$ попало в критическую область и выдвинутая гипотеза H_0 отклоняется, даже если она верна.

Ошибка первого рода состоит в том, что будет отвергнута правильная гипотеза H_0.

Вероятность совершить ошибку первого рода принято обозначать через α; ее называют ***уровнем значимости***.

Замечание. Часто уровень значимости принимают равным 0,05 или 0,01. Если, например, принят уровень значимости, равный 0,05, то это означает, что в пяти случаях из ста имеется риск допустить ошибку первого рода (отвергнуть правильную гипотезу).

2. Возможна и другая ошибка – принять гипотезу H_0, когда она неверна.

Ошибка второго рода – будет принята неправильная гипотеза H_0.

Вероятность совершить ошибку второго рода принято обозначать β; ее называют риск два.

Таблица случаев

Решение по критерию	Истина	
	H_0 верна	H_0 неверна
отклоняется	ошибка первого рода	решение верно

принимается	решение верно	ошибка второго рода

Мощностью критерия называется вероятность отклонить нулевую гипотезу, когда верна конкурирующая гипотеза

$$P_{H_1}\left(\overline{H_0}\right)=1-\beta.$$

Пусть мощность (1-β) возрастает; следовательно, уменьшается вероятность β совершить ошибку второго рода. Таким образом, чем мощность больше, тем вероятность ошибки второго рода меньше.

Итак, при заданном уровне значимости критическую область следует строить так, чтобы мощность критерия была максимальной. Выполнение этого требования должно обеспечить минимальную ошибку второго рода.

Отыскание любой из критических областей (правосторонней, левосторонней и двусторонней) сводится к нахождению критических точек $k_{кр}$.

С этой целью задаются достаточно малым уровнем значимости (обычно 0,05; 0,01). Затем ищут критическую точку исходя из условий:

– $P\left(K>k_{кр}\right)=\alpha$ (для правосторонней критической области);

– $P\left(K<k_{кр}\right)=\alpha$ (для левосторонней критической области);

– $P\left(K<k_1\right)+P\left(K>k_2\right)=\alpha$ (для двусторонней критической области).

Для каждого из критериев имеются соответствующие таблицы, по которым находят критические точки $k_{кр}$, удовлетворяющие данным требованиям.

Эмпирические и выравнивающие (теоретические) частоты

Пусть произведено n испытаний, в которых величина X приняла n_1 раз значение x_1, n_2 раз значение x_2, …, n_k раз значение x_k, причем $\sum n_i = n$.

Эмпирическими частотами называют фактические наблюдаемые частоты n_i.

Пусть имеются основания предположить, что изучаемая величина X распределена по некоторому определенному закону. Чтобы проверить, согласуется ли это предположение с данными наблюдений, вычисляют частоты наблюдаемых значений, т.е. находят теоретические частоты n'_i каждого из наблюдаемых значений в предположении, что величина X распределена по предполагаемому закону.

Выравнивающими (теоретическими) в отличие от фактических наблюдаемых эмпирических частот называют частоты n'_i, найденные теоретически (вычислением).

Опишем способ нахождения теоретических частот.

В случае **дискретного распределения** признака X генеральной совокупности выравнивающие частоты находят с помощью равенства

$$n'_i = nP_i = n\cdot P\left(X=x_i\right),$$

где n – число испытаний; P_i - вероятность наблюдаемого значения x_i, вычисленная при допущении, что X имеет предполагаемое распределение.

Итак, выравнивающая частота наблюдаемого значения x_i дискретного распределения равна произведению числа испытаний на вероятность этого наблюдаемого значения.

Важнейшим дискретным распределением является распределение Пуассона. Тогда x_i принимает значения m: 0, 1, 2 …. Вероятности P_i вычисляются по формуле Пуассона:

$$P_i = P(X = m) = \frac{\lambda^m}{m!} e^{-\lambda}.$$

Известно, что параметр λ, которым определяется распределение Пуассона, равен математическому ожиданию этого распределения. Поскольку в качестве оценки математического ожидания принимают выборочную среднюю, то и в качестве оценки λ можно принять выборочную среднюю $\bar{x}_B$.

В случае **непрерывного распределения** вероятности отдельных возможных значений равны нулю. Поэтому весь интервал возможных значений делят на k непересекающихся интервалов и вычисляют вероятности P_i попадания X в i-й частичный интервал, а затем, как и для дискретного распределения, умножают число испытаний на эти вероятности.

Итак, выравнивающие частоты непрерывного распределения находят по равенству

$$n'_i = nP_i = n \cdot P(x_i < X < x_{i+1}),$$

где n – число испытаний; P_i - вероятность попадания X в i-й частичный интервал, вычисленная при допущении, что X имеет предполагаемое распределение.

Выразив P_i через функцию распределения и плотность вероятностей, получим

$$P_i = P(x_i < X < x_{i+1}) = F(x_{i+1}) - F(x_i) = \int_{x_i}^{x_{i+1}} f(x)dx.$$

Согласно теореме о среднем

$$\int_{x_i}^{x_{i+1}} f(x)dx = (x_{i+1} - x_i) \cdot f\left(x_i^*\right),$$

где $(x_{i+1} - x_i)$ длина интервала; x_i^* – любая точка интервала $(x_i; x_{i+1})$.

В качестве точки x_i^* обычно принимают значение середины интервала $(x_i; x_{i+1})$, т.е. $x_i^* = \frac{x_i + x_{i+1}}{2}$.

Таким образом, формула для вычисления теоретических частот примет вид: $n'_i = n \cdot (x_{i+1} - x_i) \cdot f\left(x_i^*\right)$.

Методика вычисления теоретических частот нормального распределения

1) Эмпирическое распределение задано в виде последовательности интервалов одинаковой длины и соответствующих им частот

$(x_i; x_{i+1})$	$(x_1; x_2)$	$(x_2; x_3)$	…	$(x_k; x_{k+1})$

n_i	n_1	n_2	$\cdots$	n_k

Как найти теоретические частоты, если предполагается, что генеральная совокупность распределена нормально?

Воспользуемся определением выравнивающих частот непрерывного распределения:

$$n'_i = nP_i = n \cdot P(x_i < X < x_{i+1}).$$

Для нормального закона распределения

$$P(\alpha < X < \beta) = \Phi\left(\frac{\beta - a}{\sigma}\right) - \Phi\left(\frac{\alpha - a}{\sigma}\right).$$

Полагая вместо $M(X) = a$, $\bar{x}_B$ несмещенную оценку *M(X)*, а вместо $\sigma_Г - \sigma_B$, имеем

$$P(x_i < X < x_{i+1}) = \Phi\left(\frac{x_{i+1} - \bar{x}_B}{\sigma_B}\right) - \Phi\left(\frac{x_i - \bar{x}_B}{\sigma_B}\right).$$

Алгоритм вычисления теоретических частот

1. Вычислить $\bar{x}_B$ и σ_B, причем в качестве вариант принять среднее арифметическое концов интервалов $x_i^* = \frac{x_i + x_{i+1}}{2}$.

2. Пронормировать случайную величину *X*, т.е. перейти к величине $Z = \frac{\left(x_i - \bar{x}_B\right)}{\sigma_B}$, вычислив концы интервалов $(z_i; z_{i+1})$:

$$z_i = \frac{\left(x_i - \bar{x}_B\right)}{\sigma_B},\ z_{i+1} = \frac{\left(x_{i+1} - \bar{x}_B\right)}{\sigma_B},$$

причем наименьшее значение *Z*, т.е. z_1, полагают равным $-\infty$, а наибольшее, т.е. z_k, полагают равным ∞.

3. Вычислить теоретические вероятности P_i попадания *X* в интервалы $(x_i; x_{i+1})$ по равенству (*Φ(z)* – функция Лапласа)

$$P_i = \Phi(z_{i+1}) - \Phi(z_i)$$

и, наконец, найти искомые теоретические частоты $n'_i = nP_i$.

Все вычисления целесообразно внести в таблицу.

i	x_i	x_{i+1}	n_i	z_i	z_{i+1}	$\Phi(z_i)$	$\Phi(z_{i+1})$	$P_i = \Phi(z_{i+1}) - \Phi(z_i)$	$n'_i = nP_i$

2) Эмпирическое распределение задано в виде последовательности равноотстоящих вариант и соответствующих им частот. Тогда для нахождения теоретических частот используют формулу

$$n'_i = n \cdot (x_{i+1} - x_i) \cdot f\left(x_i^*\right) = n \cdot h \cdot f\left(x_i^*\right), \qquad (*)$$

где *h* – длина частичного интервала.

Запишем плотность общего нормального распределения:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{\frac{-(x-a)^2}{2\sigma^2}}.$$

При $a = 0$ и $\sigma = 1$ получим плотность нормированного распределения:

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

или, изменив обозначение аргумента,

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}e^{-\frac{u^2}{2}}.$$

Положив $u=\frac{(x-a)}{\sigma}$, имеем

$$\varphi(u)=\frac{1}{\sqrt{2\pi}}e^{\frac{-(x-a)^2}{2\sigma^2}}.$$

Сравнивая $\varphi(u)$ и *f(x)*, заключаем, что

$$f(x)=\frac{1}{\sigma}\varphi(u).$$

Если математическое ожидание *a* и среднее квадратическое отклонение σ неизвестны, то в качестве оценок этих параметров принимают соответственно выборочную среднюю $\bar{x}_B$ и выборочное среднее квадратическое отклонение σ_B.

Тогда

$$f(x)=\frac{1}{\sigma_B}\varphi(u), \qquad \text{где } u=\frac{(x-\bar{x}_B)}{\sigma_B}.$$

Вернемся к формуле (*) $n'_i = n\cdot h\cdot f\left(x_i^*\right)=n\cdot h\cdot\frac{1}{\sigma_B}\cdot\varphi(u_i)$.

Таким образом, $n'_i=\frac{n\cdot h}{\sigma_B}\cdot\varphi(u_i)$, где $u_i=\frac{(x_i-\bar{x}_B)}{\sigma_B}$.

Алгоритм нахождения теоретических частот

1. Вычислить, например методом произведений выборочную среднюю $\bar{x}_B$ и выборочное среднее квадратическое отклонение σ_B.
2. Перейти к условным вариантам $u_i=\frac{(x_i-\bar{x}_B)}{\sigma_B}$.
3. Найти значения функции Лапласа $\varphi(u_i)$.
4. Вычислить теоретические частоты по формуле $n'_i=\frac{n\cdot h}{\sigma_B}\cdot\varphi(u_i)$.

Вычисления целесообразно вносить в таблицу

i	x_i	n_i	u_i	$\varphi(u_i)$	$n'_i = \frac{n \cdot h}{\sigma_B} \cdot \varphi(u_i)$

Проверка гипотезы о нормальном распределении генеральной совокупности. Критерий согласия Пирсона

Проверка гипотезы о предлагаемом законе неизвестного распределения производится при помощи специально подобранной случайной величины – критерия согласия.

Критерием согласия называют критерий проверки гипотезы о предполагаемом законе неизвестного распределения.

Имеется несколько критериев согласия: χ^2 («хи-квадрат») Пирсона, Колмогорова, Смирнова и др. Остановимся подробнее на описании применения критерия Пирсона к проверке гипотезы о нормальном распределении генеральной совокупности (критерий аналогично применяется и для других распределений, в этом состоит его достоинство).

Критерий Пирсона служит для сравнения эмпирических и теоретических частот и отвечает на вопрос: случайно ли расхождение этих частот или оно значимо? Но критерий Пирсона, как и любой другой критерий, не доказывает справедливость гипотезы, а лишь устанавливает на принятом уровне значимости ее согласие или несогласие с данными наблюдений.

Итак, пусть по выборке объема *n* получено эмпирическое распределение

Варианты …… x_i x_1 x_2 … x_s

Эмпирические частоты ……. n_i n_1 n_2 … n_s

Допустим, что в предположении нормального распределения генеральной совокупности вычислены теоретические частоты n'_i. При уровне значимости α требуется проверить нулевую гипотезу H_0: генеральная совокупность распределена нормально.

В качестве критерия проверки нулевой гипотезы примем случайную величину

$$\chi^2 = \sum_{i=1}^{s} \frac{\left(n_i - n_i'\right)^2}{n_i'}.$$

Эта величина случайная, так как в различных опытах она принимает различные, заранее неизвестные значения. Очевидно, что чем меньше различаются эмпирические и теоретические частоты, тем меньше величина критерия χ^2 и, следовательно, разница между эмпирическим и теоретическим распределениями несущественна. Критерий согласия χ^2 характеризуется двумя параметрами: уровнем значимости α и числом степеней свободы *k*.

Число степеней свободы находят по равенству $k = s - 1 - r$, где s – число групп (частичных интервалов) выборки; r – число параметров предполагаемого распределения, которые оценены по данным выборки.

В частности, если предполагаемое распределение – нормальное, то оценивают два параметра (математическое ожидание и среднее квадратическое отклонение), поэтому $r = 2$ и число степеней свободы $k = s - 1 - r = s - 1 - 2 = s - 3$.

Если, например, предполагают, что генеральная совокупность распределена по закону Пуассона, то оценивают один параметр λ, поэтому $r = 1$ и $k = s - 2$.

Поскольку односторонний критерий более «жестко» отвергает нулевую гипотезу, чем двусторонний, построим правостороннюю критическую область исходя из требования, чтобы вероятность попадания критерия в эту область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости α:

$$P\left[\chi^2 > \chi^2_{кр}(\alpha; k)\right] = \alpha .$$

Таким образом, правосторонняя критическая область определяется неравенством $\chi^2 > \chi^2_{кр}(\alpha; k)$. Функция плотности данного распределения будет иметь вид (рис. 3.9).

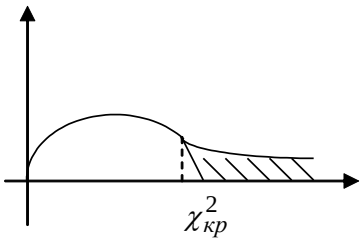


Рис. 3.9.

Существуют специальные таблицы, по которым для заданных k и α находятся соответствующие критические значения критерия $\chi^2_{кр}$ (приложение 6).

Обозначим значение критерия, вычисленное по данным наблюдений, через $\chi^2_{набл}$ и сформулируем ***алгоритм проверки нулевой гипотезы.***

1. По предполагаемому теоретическому распределению находим выравнивающие частоты n'_i.

2. Вычисляем наблюдаемое значение критерия:

$$\chi^2_{набл} = \sum_{i=1}^{s} \frac{\left(n_i - n'_i\right)^2}{n'_i} .$$

3. Находим число степеней свободы по формуле $k = s - 3$.

4. По данному значению уровня значимости α и числу степеней свободы k находим критическое значение критерия $\chi^2_{кр}$ $(\alpha; k)$.

5. Сравниваем $\chi^2_{набл}$ и $\chi^2_{кр}$. Если $\chi^2_{набл} < \chi^2_{кр}$, нет оснований отвергнуть нулевую гипотезу. Если $\chi^2_{набл} > \chi^2_{кр}$ - нулевую гипотезу отвергают.

Замечание 1. Объем выборки должен быть достаточно велик, во всяком случае, не менее 50. Каждая группа должна содержать не менее 5 – 8 вариант; малочисленные группы следует объединять в одну, суммируя частоты.

Замечание 2. Для контроля вычислений применяют формулу

$$\chi^2_{набл} = \left[\frac{\sum n_i^2}{n_i'}\right] - n\,.$$

Решение типовых задач

Задача 1. Используя критерий Пирсона, при уровне значимости α = 0,05 проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности X с эмпирическим распределением выборки объема n = 200.

x_i	5	7	9	11	13	15	17	19	21
n_i	15	26	25	30	26	21	24	20	13

Решение. 1. Вычислим $\bar{x}_B = \frac{\sum_{i=1}^{k} x_i n_i}{n} = 12{,}63$ и выборочное среднее квадратическое отклонение $\sigma_B = \sqrt{\overline{x_B^2} - \left(\bar{x}_B\right)^2} = 4{,}695$.

2. Вычислим теоретические частоты учитывая, что n = 200, h = 2, σ_B = 4,695, по формуле

$$n'_i = \frac{n \cdot h}{\sigma_B} \cdot \varphi(u_i) = \frac{200 \cdot 2}{4{,}695} \cdot \varphi\left(\frac{x_i - \bar{x}_B}{\sigma_B}\right) = 85{,}2 \cdot \varphi\left(\frac{x_i - \bar{x}_B}{\sigma_B}\right).$$

Составим расчетную таблицу (значения функции $\varphi(x)$ приведены в приложении 1).

i	x_i	$u_i = \frac{\left(x_i - \bar{x}_B\right)}{\sigma_B}$	$\varphi(u_i)$	$n'_i = \frac{n \cdot h}{\sigma_B} \cdot \varphi(u_i)$
1	5	-1,62	0,1074	9,1
2	7	-1,20	0,1942	16,5
3	9	-0,77	0,2966	25,3
4	11	-0,35	0,3752	32,0
5	13	0,08	0,3977	33,9
6	15	0,51	0,3503	29,8
7	17	0,93	0,2589	22,0
8	19	1,36	0,1582	13,5
9	21	1,78	0,0818	7,0

3. Сравним эмпирические и теоретические частоты. Составим расчетную таблицу, из которой найдем наблюдаемое значение критерия $\chi^2_{набл} = \sum_{i=1}^{s} \frac{\left(n_i - n_i'\right)^2}{n_i'}$:

i	n_i	n'_i	$\lvert n_i - n_i' \rvert$	$\left(n_i - n_i'\right)^2$	$\frac{\left(n_i - n_i'\right)^2}{n_i'}$
1	15	9,1	5,9	34,81	3,8
2	26	16,5	9,5	90,25	5,5
3	25	25,3	0,3	0,09	0,0
4	30	32,0	2,0	4,0	0,1
5	26	33,9	7,9	62,41	1,8
6	21	29,8	8,8	77,44	2,6
7	24	22,0	2,0	4,0	0,2
8	20	13,5	6,5	42,25	3,1
9	13	7,0	6,0	36,0	5,1
$\sum$	200				$\chi^2_{набл}$=22,2

По таблице критических точек распределения χ^2 (приложение 6), по уровню значимости α = 0,05 и числу степеней свободы $k = s - 3 = 9 - 3 = 6$ находим критическую точку правосторонней критической области $\chi^2_{кр}$ (0,05; 6) = 12,6.

Так как $\chi^2_{набл}$=22,2 > $\chi^2_{кр}$ = 12,6, гипотезу о нормальном распределении генеральной совокупности отвергаем. Другими словами, эмпирические и теоретические частоты различаются значимо.

Задача 2. Распределение 50 промышленных предприятий по средней численности работников характеризуются следующими данными:

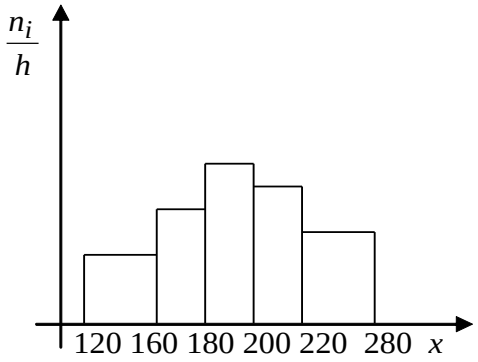
Численность работников	120-140	140-160	160-180	180-200	200-220	220-240	240-260	260-280
Число преприятий	1	4	10	14	12	6	2	1

Проверить на уровне значимости α = 0,01 гипотезу о нормальном распределении при помощи критерия Пирсона.

Решение. 1. Ввиду малочисленности частот объединяем первые два и последние три интервала. Получается таблица

$(x_i; x_{i+1})$	120-160	160-180	180-200	200-220	220-280
n_i	5	10	14	12	9

Строим гистограмму:



По виду гистограммы можно предположить, что данная случайная величина подчиняется нормальному закону распределения. Выдвинем и проверим гипотезу – H_0: исследуемая случайная величина имеет нормальный закон распределения.

2. Для вычисления теоретических частот находим $\bar{x}_B$, σ_B, n.

$$\bar{x}_B = 195{,}2;\ \sigma_B = 28{,}5;\ n = 50.$$

3. Найдем теоретические частоты $n'_i = nP_i$, где $P_i = P(x_i < X < x_{i+1})$ - вероятность того, что случайная величина попадет в интервал $(x_i; x_{i+1})$.

Так как предполагаемый закон распределения нормальный, то

$$P_i = \Phi\left(\frac{x_{i+1} - \bar{x}_B}{\sigma_B}\right) - \Phi\left(\frac{x_i - \bar{x}_B}{\sigma_B}\right),$$

где $\Phi(x)$ – функция Лапласа (приложение 3). Вычисления приведем в таблице:

i	x_i	x_{i+1}	z_i	z_{i+1}	$\Phi(z_i)$	$\Phi(z_{i+1})$	$P_i = \Phi(z_{i+1}) - \Phi(z_i)$	$n'_i = nP_i$
1	-∞	160	-∞	-1,24	-0,5	-0,3925	0,1075	5,375
2	160	180	-1,24	0,53	-0,3925	-0,2019	0,1906	9,53
3	180	200	-0,53	0,17	-0,2019	0,0675	0,2694	13,47
4	200	220	0,17	0,87	0,0675	0,3079	0,2404	12,02
5	220	+∞	0,87	+∞	0,3079	0,5	0,1921	9,605
∑							1	50

4. Сравним эмпирические и теоретические частоты, используя критерий Пирсона. Вычислим наблюдаемое значение критерия Пирсона. Для этого составим расчетную таблицу

i	n_i	n'_i	$\lvert n_i - n_i' \rvert$	$(n_i - n_i')^2$	$\frac{(n_i - n_i')^2}{n_i'}$
1	5	5,375	0,375	0,1406	0,0262
2	10	9,53	0,47	0,2209	0,0223

3	14	13,47	0,53	0,2809	0,0209
4	12	12,02	0,02	0,0004	0,0
5	9	9,605	0,605	0,366	0,0381
∑	50	50			$\chi^2_{набл}$=0,1075

По таблице критических точек распределения χ^2 (приложение 6), по уровню значимости α = 0,01 и числу степеней свободы $k = s - 3 = 5 - 3 = 2$ (s – число интервалов) находим критическую точку правосторонней критической области $\chi^2_{кр}$ (0,01; 2) = 9,2.

Сравним $\chi^2_{набл}$ и $\chi^2_{кр}$. Так как $\chi^2_{набл}$=0,1075 < $\chi^2_{кр}$ = 9,2, нет оснований отклонить гипотезу о нормальном распределении генеральной совокупности. Другими словами, эмпирические и теоретические частоты различаются незначимо.

Задачи (91 – 100)

В задачах 91 – 100 даны эмпирические значения случайной величины X. Требуется:

1. Выдвинуть гипотезу о виде распределения.
2. Проверить гипотезу с помощью критерия Пирсона при заданном уровне значимости α.

За значения параметров a и σ принять среднюю выборочную и среднее выборочное квадратическое отклонение, вычисленные по эмпирическим данным.

91. В таблице дано распределение среднегодовой стоимости основных фондов (млн руб.) по 50 предприятиям отрасли.

x_i	100,5	111,2	121,5	132,0	142,5	153,0	163,5
n_i	4	9	18	8	5	4	2

α = 0,05.

92. По данным, полученным от 50 фермерских хозяйств одного из регионов, составлено распределение численности работников.

x_i	1-5	5-9	9-13	13-17	17-21	21-25
n_i	6	10	17	12	4	1

α = 0,01.

93. Распределение 60 магазинов по величине товарооборота (млн руб.) характеризуется следующими данными:

x_i	1,0-1,5	1,5-2,0	2,0-2,5	2,5-3,0	3,0-3,5
n_i	5	11	23	13	8

α = 0,05.

94. Результаты анализа темпов роста стоимости акций 50 компаний (%) имеют следующее распределение:

x_i	102-104	104-106	106-108	108-110	110-112
n_i	5	10	15	12	8

α = 0,05.

95. Распределение 50 туристических фирм по средней численности работников характеризуется следующими данными:

x_i	0-5	5-10	10-15	15-20	20-25	25-30
n_i	4	7	11	21	5	2

α = 0,025.

96. В таблице дано распределение величины дохода торговых предприятий за год (тыс. руб.):

x_i	20-24	24-28	28-32	32-36	36-40
n_i	10	21	30	17	12

α = 0,01.

97. Распределение стоимости покупок (руб.) 50 случайно выбранных покупателей характеризуется следующими данными:

x_i	10-20	20-30	30-40	40-50	50-60	60-70	70-80	80-90	90-100	100-110
n_i	1	3	4	6	11	10	7	5	2	1

α = 0,01.

98. Результаты анализа 100 промышленных предприятий по возрастной структуре производственного оборудования характеризуются следующими данными:

x_i	0-5	5-10	10-15	15-20	20-25	25-30	30-35
n_i	4	15	20	26	19	14	2

α = 0,025.

99. Распределение 50 промышленных предприятий по уровню механизации труда (%) характеризуется следующими данными:

x_i	5-15	15-25	25-35	35-45	45-55	55-65
n_i	6	5	10	13	9	7

α = 0,01.

100. В таблице дано распределение расходов на рекламу у предприятий в долях от дохода:

x_i	0,2-0,4	0,4-0,6	0,6-0,8	0,8-1	1-1,2	1,2-1,4
n_i	5	17	23	16	7	2

α = 0,01.

Вопросы для самоконтроля

1. Несмещенной оценкой $\overline{X}_Г$ ***является***

101. $M\left(\overline{X}\right)$. ***102.*** $\overline{X}$ ***103.*** $\overline{X}_B$. ***104*** $\overline{X^2}$.

2. Дано распределение выборки

x_i	***10***	***15***	***20***
n_i	***2***	***5***	***3***

Найти $\overline{X}_B$

201. 15,5. ***202.*** 1,55. ***203.*** 15. ***204.*** 155.

3. Выборочный коэффициент корреляции $r_B = 0{,}85$***. Оценить тесноту связи.***

301. умеренная; ***302.*** высокая; ***303.*** заметная; ***304.*** весьма высокая.

4. Статистическая оценка не должна удовлетворять требованию

401. несмещенность. ***402.*** эффективность.
403. состоятельность. ***404.*** неотрицательность.

5. Вероятность, с которой осуществляется неравенство $\left|\theta - \theta^*\right| < \delta$***, называют***

501. точностью; ***502.*** оценкой; ***503.*** надежностью; ***504.*** эффектом.

6. Известно, что уровень значимости $\alpha = 0{,}05$***, число степеней свободы*** $k = 13$***. Каково критическое значение критерия*** $\chi^2_{кр}$***?***

601. 23,7. ***602.*** 22,4. ***603.*** 26,1. ***604.*** 24,7.

7. Согласно статистическому критерию Пирсона, нулевую гипотезу отвергают, если

701. $\chi^2_{набл} > \chi^2_{кр}$. ***702.*** $\chi^2_{набл} \neq \chi^2_{кр}$. ***703.*** $\chi^2_{набл} = \chi^2_{кр}$. ***704.*** $\chi^2_{набл} < \chi^2_{кр}$.

8. Зависимость, при которой изменение одной из величин влечет изменение среднего значения другой величины называется

801. статистической; ***802.*** функциональной;
803. корреляционной; ***804.*** многомерной.

9. В качестве критерия согласия проверки гипотезы о нормальном распределении генеральной совокупности применяется случайная величина $\chi^2 = \sum_{i=1}^{s} \frac{\left(n_i - n_i'\right)^2}{n_i'}$***, где*** n'_i ***- …***

901. стандартизированные частоты; ***902.*** эмпирические частоты;
903. нормированные частоты; ***904.*** теоретические частоты.

10. Найти параметр t в доверительном интервале $\overline{X} - t\frac{\sigma}{\sqrt{n}} < a < \overline{X} + t\frac{\sigma}{\sqrt{n}}$***, если надежность γ = 0,95.***

1001. 0,47. ***1002.*** 2,04. ***1003.*** 1,96. ***1004.*** 0,52.