МИНИСТЕРСТВО НАУКИ И ВЫСШЕГО ОБРАЗОВАНИЯ
РОССИЙСКОЙ ФЕДЕРАЦИИ
ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ
ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ
"МУРМАНСКИЙ АРКТИЧЕСКИЙ УНИВЕРСИТЕТ"

ТЕОРИЯ ВЕРОЯТНОСТЕЙ И МАТЕМАТИЧЕСКАЯ СТАТИСТИКА
Часть 2
МАТЕМАТИЧЕСКАЯ СТАТИСТИКА

Рекомендовано учебно-методическим советом университета
в качестве учебного пособия
для обучающихся по направлениям подготовки
01.03.02 "Прикладная математика и информатика",
44.03.05 "Педагогическое образования",
09.03.01 "Информатика и вычислительная техника"

Учебное электронное издание

Мурманск
Издательство МАУ
2024

16+

УДК 519.2(075.8)
ББК 22.17я73
Т 33

Рецензенты:
А. В. Немыкин, заведующий кафедрой экономики филиала РАНХиГС г. Мурманска, канд. экон. наук;
МБУ ДПО г. Мурманска "Городской информационно-методический центр работников образования"

Т 33 **Теория вероятностей и математическая статистика. Часть 2. Математическая статистика**: учеб. пособие для студентов высших учебных заведений / авт.-сост. В. В. Левитес; М-во науки и высш. образования Рос. Федерации, Мурм. аркт. ун-т. – Мурманск : Изд-во МАУ, 2024. – 1 опт. компакт-диск (CD-ROM). – Систем. требования: PC не ниже класса Pentiuum II 128 MbRAM ; Windows 9X–10 ; свободное место на HDD 131 Mb ; привод для компакт-дисков CD-ROM 2-х и выше. – Загл. с титул. экрана. – Текст : электронный.

ISBN 978-5-907368-82-8 (общ.)
ISBN 978-5-907905-05-4 (Ч. 2)

Учебное пособие предназначено для помощи студентам в их самостоятельной работе при изучении раздела "Математическая статистика". Пособие содержит теоретический материал из общего курса математической статистики, а также включает задачи для самостоятельной работы. Теоретический материал проиллюстрирован примерами. В пособии приведены таблицы значений функций, необходимые для решения задач (Приложения), а также список рекомендуемой литературы.

The manual is intended to help students in their independent work when studying the section "Mathematical Statistics". The manual contains theoretical material from the general course of mathematical statistics, and also includes tasks for independent work. The theoretical material is illustrated with examples. The manual contains tables of function values necessary for solving problems (Appendices), as well as a list of recommended literature.

Учебное электронное издание
Минимальные системные требования:
PC не ниже класса PentiuumII 128 MbRAM;
свободное место на HDD 131 Mb;
привод для компакт-дисков CD-ROM 2x и выше.

Учебное электронное издание

Минимальные системные требования:
PC не ниже класса PentiuumII 128 MbRAM;
свободное место на HDD 131 Mb;
привод для компакт-дисков CD-ROM 2x и выше.

Вера Владимировна Левитес

Рецензенты:
А. В. Немыкин, заведующий кафедрой экономики филиала РАНХиГС г. Мурманска, канд. экон. наук;
МБУ ДПО г. Мурманска "Городской информационно-методический центр работников образования"

Компьютерная верстка Г. Г. Недоступ

Подписано к использованию 15.07.2024
Объём издания 1,89 Мб
Тираж 30 экз.

ФГАОУ ВО "Мурманский арктический университет"
183010, г. Мурманск, ул. Спортивная, 13.
Телефон (8152) 21-38-01
E-mail: office@mauniver.ru
https://www.mauniver.ru

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Массовые случайные явления подчиняются различным закономерностям, которые изучаются методами теории вероятностей статистических данных. Для понимания характера изучаемой случайной величины нужно знать ее закон распределения. Определение законов распределения рассматриваемых величин и оценка значений параметров распределения на основании наблюденных значений – задача математической статистики.

Еще одной задачей математической статистики является создание методов обработки и анализа статистического материала в зависимости от целей исследования.

Методическое пособие предназначено для помощи студентам в их самостоятельной работе при изучении раздела "Математическая статистика".

Пособие содержит теоретический материал из общего курса математической статистики, а также включает задачи для самостоятельной работы. Теоретический материал проиллюстрирован примерами. В пособии приведены таблицы значений функций, необходимые для решения задач (Приложения), а также список рекомендуемой литературы.

Пособие может быть использовано в качестве основной литературы для проведения лекций и практических занятий для различных направлений подготовки, таких как 01.03.02 Прикладная математика и информатика, 09.03.01 Информатика и вычислительная техника, 44.03.05 Педагогическое образование.

1. ОСНОВНЫЕ ПОНЯТИЯ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Математическая статистика занимается установлением закономерностей, которым подчинены массовые случайные явления, на основе обработки статистических данных, полученных в результате наблюдений. Двумя основными задачами математической статистики являются:

1) определение способов сбора и группировки этих статистических данных;

2) разработка методов анализа полученных данных в зависимости от целей исследования, к которым относятся:

а) оценка неизвестной вероятности события; оценка неизвестной функции распределения; оценка параметров распределения, вид которого известен; оценка зависимости от других случайных величин и т. д.;

б) проверка статистических гипотез о виде неизвестного распределения или о значениях параметров известного распределения.

1.1. Выборки и их виды

Изучение закономерностей объектов достаточно большой совокупности методами математической статистики основано на использовании статистических данных для некоторой конечной части рассматриваемых объектов.

Для решения этих задач необходимо выбрать из большой совокупности однородных объектов ограниченное количество объектов, по результатам изучения которых, можно сделать прогноз относительно исследуемого признака этих объектов.

Признак – это объективная характеристика единицы статистической совокупности, характерная черта или свойство, которое может быть определено или измерено.

Признаки подразделяются на количественные и качественные, а последние, в свою очередь, на альтернативные, атрибутивные и порядковые.

Количественный признак – отдельные варианты, которого имеют числовое выражение и отражают размеры, масштабы изучаемого объекта или явления.

Качественный признак выражается смысловым понятием (отдельные значения которого выражаются в виде понятий, наименований).

Альтернативный признак – имеет только два варианта значений. **Атрибутивные** признаки имеют более двух вариантов, которые при этом выражаются в виде понятий или наименований.

Пример: район проживания, вид продукции, специальность работника, цвет товара.

Такие признаки имеют место в различных областях исследования, но в большей степени они характерны для информации, с которой работают маркетологи, социологи, психологи.

Порядковые признаки – имеют несколько ранжированных, т. е. упорядоченных по возрастанию или убыванию, качественных вариантов.

Пример: уровень образования (начальное, общее среднее и т. д.), уровень квалификации, воинское звание, различного рода рейтинги.

Отдельные варианты порядкового признака трудно соизмерить количественно.

Вариант – это возможное значение, которое может принимать признак.

Вариация – это колеблемость, изменение величины признака в статистической совокупности, т. е. принятие единицами совокупности или их группами разных значений признаков. Приведенные выше примеры показывают, что изучаемые статистикой признаки как правило подвержены вариации.

Объектом любого статистического исследования является статистическая совокупность.

Допустим, у нас есть некоторая совокупность однородных объектов и нас интересует некоторый количественный или качественный признак, характеризующий эти объекты, например, размер деталей, в магазине – вес расфасованных продуктов. Данный признак мы будем интерпретировать как случайную величину, значение которой меняется от объекта к объекту. Иногда проводят сплошное обследование – обследуют каждый объект совокупности относительно признака, которым интересуются. Но не всегда это возможно. Обычно из всей совокупности объектов случайным образом отбирают ограниченное число объектов, которые и подвергают изучению.

Статистическая совокупность – это группа относительно однородных элементов, взятых вместе в конкретных границах пространства и времени и обладающих признаками сходства и различия.

Единица совокупности – индивидуальный составной элемент статистической совокупности, являющийся носителем изучаемых признаков.

Объем совокупности – общее число единиц, образующих статистическую совокупность, следует отличать от объема признака.

Однородной является совокупность, единицы которой близки между собой по значениям признаков, существенных для данного исследования, или же они относятся к одному и тому же типу. Многие методы и приемы статистического исследования применимы лишь к однородным совокупностям.

Различают два вида статистической совокупности: генеральную и выборочную.

Генеральная совокупность – совокупность, состоящая из всех единиц наблюдения, которые могут быть отнесены к ней в соответствии с целью исследования. При изучении общественного здоровья генеральная совокупность часто рассматривается в пределах конкретных территориальных границ или может ограничиваться другими признаками (пол, возраст и др.) в зависимости от цели исследования. Генеральная совокупность – все множество имеющихся объектов

Выборочная совокупность (выборка) – часть генеральной совокупности, отобранная специальным (выборочным) методом и предназначенная для характеристики генеральной совокупности.

Выборку можно рассматривать как некоторый эмпирический аналог генеральной совокупности.

Объем генеральной совокупности N и объем выборки n – число объектов в рассматриваемой совокупности.

Пример.

Количество зарегистрированных малых предприятий торговли продуктами питания в городе Мурманске равно 2436. Для исследования предприятий по объему товарооборота взято 136 предприятий. В данном случае $N = 2436$ – объем генеральной совокупности (все мыслимые предприятия данной категории), $n = 136$ – объем выборки из генеральной совокупности.

Различают повторную и бесповторную выборки (см. рис. 1).

Важнейшей задачей выборочного метода является оценка параметров (характеристик) генеральной совокупности по данным выборки. По некоторой части генеральной совокупности (т. е. по выборке) можно сделать вывод о ее свойствах в целом.

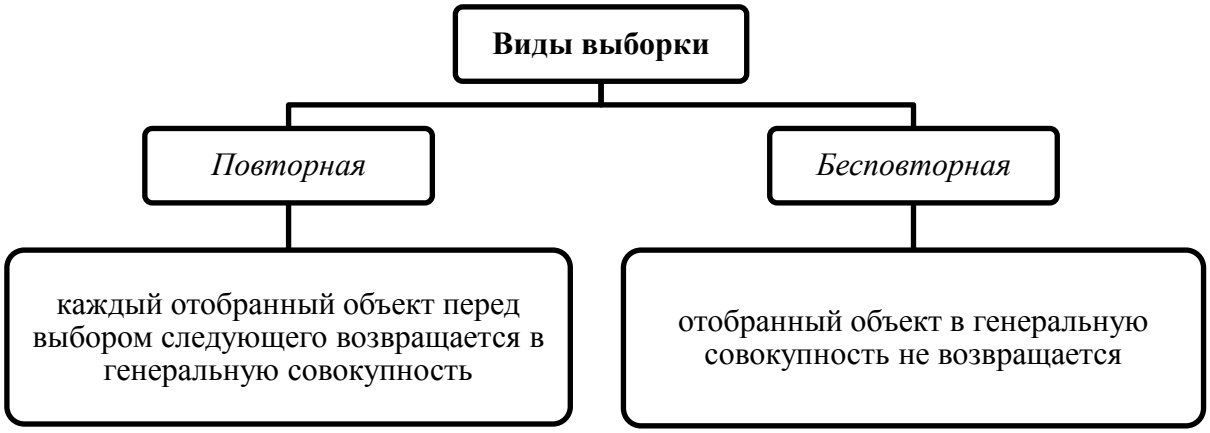


Рис. 1. Виды выборок

Теоретическую основу применимости выборочного метода составляет закон больших чисел, согласно которому при неограниченном увеличении выборки практически достоверно, что случайные выборочные характеристики как угодно близко приближаются по вероятности к ограниченным параметрам генеральной совокупности.

Выборочная совокупность должна быть репрезентативной (представительной), точно и полно отражать явление, т. е. давать такое же представление о явлении, как если бы изучалась вся генеральная совокупность.

Репрезентативность (представительность) выборки – это способность выборки воспроизводить определенные характеристики генеральной совокупности в пределах допустимых погрешностей.

В репрезентативной выборке все основные признаки генеральной совокупности, из которой извлечена данная выборка, представлены приблизительно в той же пропорции или с той же частотой, с которой данный признак выступает в этой генеральной совокупности.

Выборку называют репрезентативной, если результат измерения определенного параметра для данной выборки совпадает с учетом допустимой погрешности с известным результатом измерения генеральной совокупности. Если выборочное измерение отклоняется от известного параметра генеральной совокупности больше выбранного уровня погрешности, то такая выборка считается нерепрезентативной.

Проверка репрезентативности осуществляется посредством сравнения отдельных параметров выборки и генеральной совокупности. Распространенным заблуждением является существование репрезентативных выборок "вообще".

Репрезентативность или нерепрезентативность выборки может быть установлена исключительно в отношении отдельных переменных. Более того, одна и та же выборка может быть репрезентативна по одним параметрам и нерепрезентативна – по другим.

Принято считать, что при $n > 60$ выборка большая, или репрезентативная, а при $n < 60$ – малая. Такое деление выборки на большую и малую условно.

Понятие **репрезентативная выборка** не всегда можно связать с ее объемом n. Чаще это зависит от реально исследуемого объекта или явления, объема генеральной совокупности, трудоемкости и стоимости получения наблюдений или измерений для формирования выборки. Возможны ситуации, когда генеральная совокупность мала. Например, исследуется время наработки до отказа уникального оборудования, когда в эксплуатации находится заведомо малое количество его экземпляров. Доступного для исследования оборудования может быть еще меньше. Поэтому выборка объемом n, близким к объему генеральной совокупности N, может считаться репрезентативной и одновременно малой $(n < 60)$.

Для обеспечения репрезентативности выборочная совокупность должна отвечать следующим требованиям:

1) быть подобной генеральной совокупности, обладать основными чертами ее, т. е. в отобранной части должны быть представлены все элементы в таком же соотношении, как и в генеральной

2) каждый элемент выборки x_i выбран случайно;

3) все x_i имеют одинаковую вероятность попасть в выборку;

4) объем выборки n должен быть настолько велик, насколько позволяет решать задачу с требуемым качеством, т. е. выборка должна быть репрезентативной.

Достоинства выборочного метода:

– позволяет существенно экономить затраты ресурсов;

– является единственно возможным в случае бесконечной генеральной совокупности;

– при тех же затратах ресурсов дает возможность проведения углубленного исследования за счет расширения программы исследования;

– позволяет снизить ошибки регистрации.

Недостатки выборочного метода:

– ошибки исследования, называемые ошибками репрезентативности.

Однако неизбежные ошибки могут быть и заранее оценены с помощью правильной организации выборки и сведены к практически незначительным величинам.

Статистика позволяет с помощью специальных формул или готовых таблиц рассчитать необходимое число наблюдений в выборочной совокупности и располагает **способами формирования выборки**: случайный, механический, типологический, гнездовой, направленный отбор.

Случайный отбор – это отбор по жребию, по начальной букве фамилии или дню рождения, по таблице случайных чисел.

Механический отбор – это отбор из генеральной совокупности каждой *n*-й единицы наблюдения (каждая 5-я, 10-я и т. д.) без учета типичности или важности отдельных частей явления.

Типологический отбор предполагает разбивку изучаемого материала на ряд однотипных качественных групп, из которых далее отбираются единицы для наблюдения.

Гнездовой (серийный) отбор – это отбор из всей совокупности групп, называемых гнездами. Затем в этих гнездах единицы наблюдения изучаются сплошным методом или выборочно.

Направленный отбор наиболее часто применяется в биологических экспериментах, реже – в социально-гигиенических исследованиях; использование этого метода позволяет выявить влияние неизвестных факторов при устранении влияния известных.

Статистическое исследование независимо от его масштабов и целей всегда завершается расчетом и анализом различных по виду и форме выражения статистических показателей.

Статистический показатель представляет собой количественную характеристику социально-экономических явлений и процессов в условиях качественной определенности. Качественная определенность показателя заключается в том, что он непосредственно связан с внутренним содержанием изучаемого явления или процесса, его сущностью.

Установление статистических закономерностей, присущих массовым случайным явлениям, основано на изучении статистических данных – сведений о том, какие значения принял в результате наблюдений интересующий нас признак *X*.

1.2. Дискретный статистический ряд

Вариационным рядом называется последовательность всех элементов выборки, расположенных в неубывающем порядке. Одинаковые элементы повторяются.

Запись вариационного ряда: $x_1, x_2, \ldots, x_n$.

Элементы вариационного ряда называют его **вариантами** или порядковыми статистиками.

Пример.

Студенты получили следующие баллы по тесту: 11, 8, 9, 10, 8, 6, 7, 7, 9, 11, 10, 6, 5, 11, 10. Записать вариационный ряд.

Решение:

Расположим данные в порядке возрастания: 5, 6, 6, 7, 7, 8, 8, 9, 9, 10, 10, 10, 11, 11, 11 – это *вариационный ряд*.

Вариационный ряд называется дискретным, если любые его варианты отличаются на конечную постоянную величину, и называется непрерывным (или интервальным), если его варианты могут отличаться друг от друга на сколь угодно малую величину.

Дискретным статистическим рядом называется последовательность различных вариант x_i с указанием частот повторения элементов.

Пусть интересующая нас случайная величина X принимает в выборке значение $x_1 - n_1$ раз, $x_2 - n_2$ раз, …, $x_к - n_к$ раз, причем $\sum_{i=1}^{k} n_i = n$, где n – объем выборки. Наблюдаемые значения случайной величины $x_1, x_2, \ldots, x_к$ называют вариантами, а $n_1, n_2, \ldots, n_к$ – частотами. Перечень вариант и соответствующих им частот или относительных частот называется **статистическим рядом**.

Если разделить каждую частоту на объем выборки, то получим относительные частоты $w_i = \frac{n_i}{n}$ (*частости*). Очевидно, что сумма частот равна объему выборки (выборочной совокупности) n, а сумма относительных частот (частостей) равна единице: $\sum_{i=1}^{k} w_i = \sum_{i=1}^{k} \frac{n_i}{n} = 1$

Дискретный статистический ряд можно записать в виде таблицы:

x_i	x_1	x_2	…	x_k
n_i	n_1	n_2	…	n_k
w_i	w_1	w_2	…	w_k

Пример.

При проведении 20 серий из 10 бросков игральной кости число выпадений шести очков оказалось равным 1,1,4,0,1,2,1,2,2,0,5,3,3,1,0,2,2,3,4,1. Составить вариационный и статистический ряды.

Решение.

Составим вариационный ряд:

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

Статистический ряд для абсолютных и относительных частот имеет вид:

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

Если в статистическом распределении вместо частот (относительных частот) указать накопленные частоты (относительные накопленные частоты), то такой ряд распределения называют *кумулятивным*.

Накопленная частота представляет собой сумму частот всех значений, от x_1 до x_i. $F_i = \sum_{j=1}^{i} n_j$. По накопленной частоте можно определить, для какой части выборки значения переменной X не превосходят значения x_i.

1.3. Группированный статистический ряд

Если исследуется некоторый непрерывный признак или число его значений ее велико, то вариационный ряд может состоять из очень большого количества чисел. В этом случае удобнее использовать **группированную выборку или интервальный статистический ряд**, под которым понимают упорядоченную совокупность интервалов варьирования значений случайной величины с соответствующими частотами или частостями попаданий в каждый из них значений случайной величины.

Для получения группированной выборки интервал, в котором заключены все наблюдаемые значения признака, разбивают на несколько равных частичных интервалов длиной h, а затем находят для каждого частичного интервала n_i – сумму частот вариант, попавших в i-й интервал. Составленная по этим результатам таблица называется **группированным статистическим рядом**:

Номера интервалов	1	2	…	k
Границы интервалов	$[a, a + h)$	$[a + h, a + 2h)$	…	$[b – h, b]$
Сумма частот вариант, попавших в интервал	n_1	n_2	…	n_k

Построение интервального ряда распределения включает в себя несколько этапов.

1) Определить минимальное и максимальное значение вариант и рассчитываем **размах** вариационного ряда по формуле: $R=x_{max}-x_{min}$

2) Рассчитать число классов N.

Наиболее часто используемыми формулами для определения числа интервалов являются две: $N=\sqrt{n}$ и N=1+3,322·lg$n$ (формула Стерджесса)

Эти формулы дают схожую оценку числа интервалов при общей численности совокупности примерно до 50 единиц. При большей совокупности обнаруживаются большие различия. Например, при n = 100, по первой формуле число интервалов равно 10, а по второй – 7, при n = 1000 соответственно 32 и 10, поэтому предпочтение следует отдавать формуле Стерджесса.

Можно определять число интервалов, в зависимости от объема выборки, с помощью таблицы.

Объем выборки	25-40	40-60	60-100	100-200	более 200
Число интервалов	5-6	6-8	7-10	8-12	10-15

3) Рассчитать интервал каждого класса.

Любой интервал содержит нижнюю и верхнюю границы Шаг интервала – это разницу между этими границами. Шаг для всех интервалов должен быть одинаковым.

Для расчета шага интервала используется формула: $h=\frac{R}{N}$.

Если при изучении ранжированного ряда обнаружится, что максимальное или минимальное (или даже несколько значений) сильно отличаются от остальных, то при расчете шага интервала следует использовать соответственно не максимальное, а предшествующее ему значение, не минимальное, а следующее в ранжированном ряду значение признака. В противном случае может получиться, что в одном-двух интервалах будут сосредоточены все наблюдения.

Шаг интервала обычно рассчитывают с той же точностью, с какой представлены значения признака в изучаемой совокупности. Если при расчете шага интервала требуется округление до заданной точности, то **округление** производят всегда в **большую сторону.**

4) Составить таблицу границ классов.

Первый интервал в качестве нижней границы имеет x_{min}, в качестве верхней $x_{min} + h$; второй интервал в качестве нижней имеет верхнюю границу первого интервала ,то есть $x_{min} + h$, для получения верхней границы этого интервала надо вновь прибавить шаг интервала , то есть $x_{min} + 2h$ и так далее.

Если при определении шага интервала пришлось отказаться от x_{min}, то в первом интервале сразу находится верхняя граница, для чего к значению, которое использовалось при расчете шага интервала следует прибавить шаг интервала, нижняя граница первого интервала не обозначается. Сам интервал будет открыт снизу. Если при расчете шага интервала пришлось отказаться от максимального значения, для того, чтобы и это значение присутствовало в интервальном ряду, открытым сверху делают последний интервал.

Можно за начало первого интервала брать величину $x_{\min} - h/2$, а конец последнего должен быть больше x_{max}.

После составления границ классов необходимо обязательно проверить, что максимальное значение выборки попало в последний интервал.

5) Подсчитать сколько единиц попало в каждый интервал.

Иногда интервальный статистический ряд, для простоты исследований, условно заменяют дискретным. В этом случае серединное значение i-го интервала принимают за вариант x_i, а соответствующую интервальную частоту n_i – за частоту этого варианта.

Пример:

Составить группированный статистический ряд 20 исследуемых по показателям результатов тестирования прыжка в высоту, если данные выборки таковы: x_i, см ~ 185, 170, 190, 170, 190, 178, 188, 175, 192, 178, 176, 180, 185, 176, 180, 192, 190, 190, 192, 194.

Решение:

1) Определяем минимальное и максимальное значение вариант и рассчитываем размах вариационного ряда:

$x_{max} = 194$; $x_{min} = 170$; $R = 194 - 170 = 24$ см

2) Рассчитываем число классов по формуле Стерджесса:

$N = 1 + 3{,}322\lg 20 = 5{,}30631 \approx 5$

3) Рассчитываем интервал каждого класса: $h = \dfrac{24}{5} = 4{,}8$

Так как в выборке все значения целые возьмем $h = 5$.

4) Составляем таблицу границ классов.

Номер интервала	1	2	3	4	5
интервал x_i - x_{i+1}	170-175	175-180	180-185	185-190	190-195
Частота класса n_i	2	5	2	3	8

Группированная форма представления случайной величины не содержит информации о каждом элементе выборки. При этом часто в качестве значения случайной величины на каждом интервале принимается его середина.

От негруппированной выборки всегда можно перейти к группированной, но не наоборот. Необходимо помнить, что переход к группированной форме представления выборки сопряжен с потерей информации об исследуемом объекте, процессе или явлении.

1.4. Графическое представление вариациионого ряда

Для наглядного представления о поведении исследуемой случайной величины в выборке можно строить различные графики.

Полигон частот – ломаная, отрезки которой соединяют точки с координатами (x_1, n_1), (x_2, n_2), …, (x_k, n_k), где x_i откладываются на оси абсцисс, а n_i – на оси ординат (см. рисунок 2). Если на оси ординат откладывать не абсолютные (n_i), а относительные (w_i) частоты, то получим **полигон относительных частот** (см. рисунок 2).

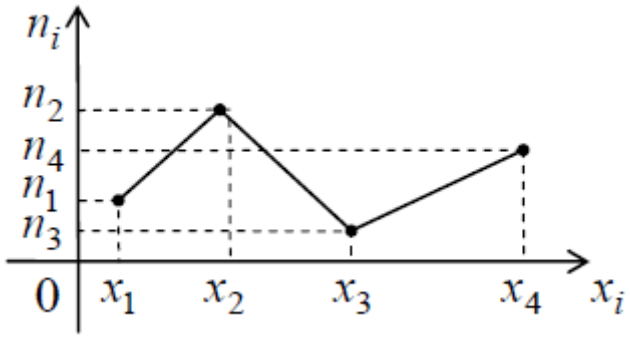


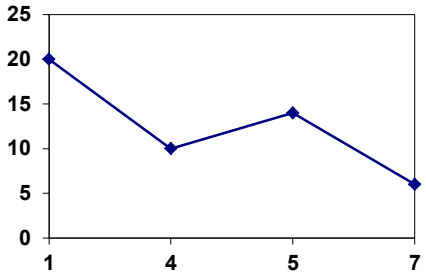
Рис. 2. Полигон частот

Пример:

Построить полигон частот по данным выборки

x_i	1	4	5	7
n_i	20	10	14	6

Решение.



В случае непрерывного признака X целесообразно строить различные гистограммы.

Гистограмма – ступенчатая фигура, состоящая из прямоугольников, основаниями которых служат частичные интервалы длиной h, а высотами – отрезки длиной n_i/h (гистограмма частот) или w_i/h (гистограмма относительных частот). В первом случае площадь гистограммы равна объему выборки, во втором – единице.

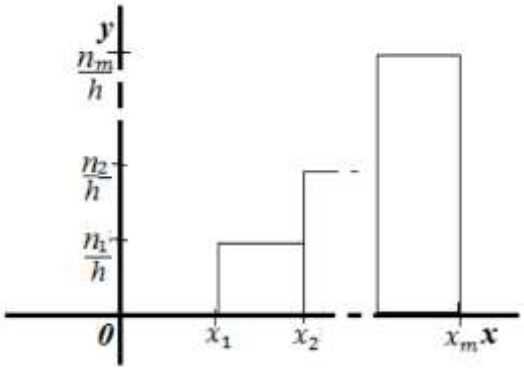
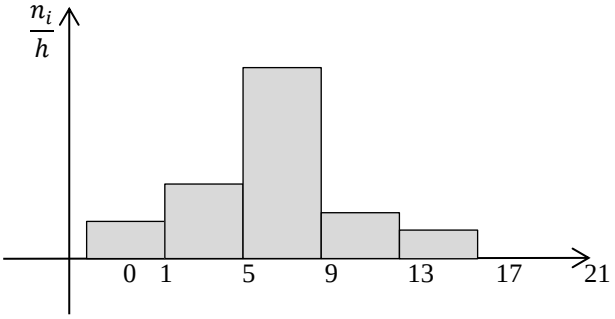


Рис. 3. Гистограмма частот

Пример:

$x_i - x_{i+1}$	1-5	5-9	9-13	13-17	17-21
n_i	10	20	50	12	8
n_i/h	2,5	5	12,5	3	2

Решение.



Кумулятивные ряды графически изображают в виде **кумуляты**. Для ее построения на оси абсцисс откладывают варианты признака или интервалы, а на оси ординат – накопленные частоты $F(x)$ или относительные накопленные частоты, а затем точки с координатами $(x_i; F(x_i))$ или $(x_i; F^*(x_i))$ соединяют отрезками прямой. В теории вероятностей кумуляте соответствует график интегральной функции распределения $F(x)$.

Пример.

Имеется распределение 80 предприятий по числу работающих на них (чел.). Найти накопленные частоты $F(x_i)$ и построить кумуляту.

i	1	2	3	4	5	6	7
x_i	150	250	350	450	550	650	750
n_i	1	3	7	30	19	15	5
$F(x_i)$	1	4	11	41	60	75	80

На рисунке 4 показана кумулята распределения предприятий по числу работающих (чел.).

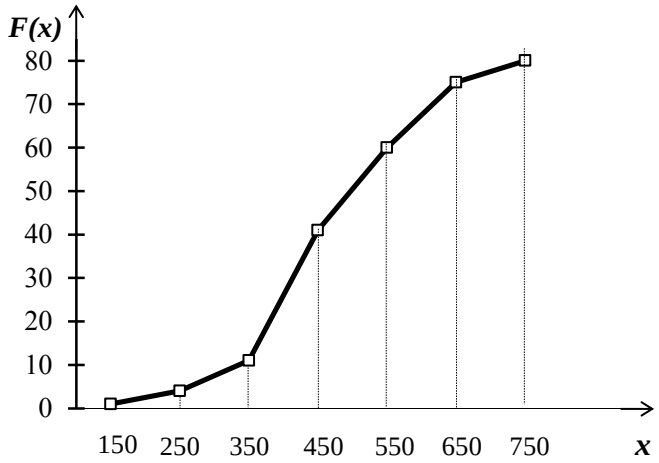


Рис. 4. Кумулята распределения

Графическое представление результатов измерений не только существенно облегчает анализ и выявление скрытых закономерностей, но и позволяет правильно выбрать последующие статистические характеристики и методы.

Если гистограмма и полигон по своему виду близки к виду графика нормального распределения, то группа однородна.

Если графики низкие и растянутые, то группа возможно однородна, но не компактна.

Если графики имеют две и более вершины, то группа неоднородна по данному признаку и ее необходимо разбить на группы.

1.5. Эмпирическая функция распределения

Вариационный ряд является статистическим аналогом (реализацией) распределения признака (случайной величины *X*). В этом смысле полигон или гистограмма аналогичен кривой распределения, а эмпирическая функция распределения – функции распределения случайной величины *X*.

Выборочной (эмпирической) функцией распределения называют функцию $F^*(x)$, определяющую для каждого значения *x* относительную частоту события $X < x$. Таким образом, $F^*(x) = \frac{n_x}{n}$, где n_x – число вариант, меньших *x*, *n* – объем выборки.

Из определения эмпирической функции распределения видно, что ее свойства совпадают со свойствами *F*(*x*), а именно:

1) $0 \le F^*(x) \le 1$.

2) $F^*(x)$ – неубывающая функция.

3) Если x_1 – наименьшая варианта, то $F^*(x) = 0$ при $x \le x_1$; если $x_к$ – наибольшая варианта, то $F^*(x) = 1$ при $x > x_к$.

Пример.

Построить эмпирическую функцию распределения для статистического ряда:

x_i	1	4	6
n_i	10	15	25

Решение.

$n = 50$

При $x \le 1$ $F^*(x) = 0$

При $1 < x \le 4$, $x_1 = 1$ наблюдается 10 раз $F^*(x) = \frac{10}{50} = 0{,}2$

При $4 < x \le 6$ $x_1 = 1$, $x_2 = 4$ наблюдается 25 раз $F^*(x) = \frac{25}{50} = 0{,}5$.

Получаем:

$$F^*(x) = \begin{cases} 0, x \le 1 \\ 0{,}2; 1 < x \le 4 \\ 0{,}5; 4 < x \le 6 \\ 1, x > 6 \end{cases}$$

Вопросы для самоконтроля

1. Каковы основные задачи математической статистики?
2. Что называется генеральной и выборочной совокупностями для исследуемой случайной величины?
3. В чем сущность выборочного метода?
4. Как получают повторную и бесповторную выборки?
5. Какая выборка называется репрезентативной, однородной?
6. Что такое частота появления варианты в выборке?
7. Как получают относительную частоту варианты в выборке?
8. Как получают вариационный ряд распределения?
9. Что такое группированный статистический ряд?
10. Как построить по данной выборке дискретный и интервальный сгруппированные статистические ряды?
11. Что такое полигон частот?
12. Как построить многоугольник распределения относительных частот?
13. Как построить гистограмму распределения плотностей относительных частот?

Задачи для самостоятельного решения

1. Выборка задана в виде распределения частот. Найти распределение относительных частот.

x_i	5	6	7
n_i	1	3	6

2. Найти эмпирическую функцию по данному распределению выборки.

а)

x_i	2	5	7	8
n_i	1	3	2	4

б)

x_i	4	7	8
n_i	5	2	3

3. Построить полигон частот по данному распределению выборки.

а)

x_i	2	3	5	6
n_i	10	15	5	20

б)

x_i	15	20	25	30	35
n_i	10	15	30	20	25

4. Построить полигон относительных частот по данному распределению выборки.

а)

x_i	2	3	5	6
n_i	10	15	5	20

б)

x_i	15	20	25	30	35
n_i	10	15	30	20	25

5. Построить гистограмму относительных частот по данному распределению выборки.

$x_i - x_{i+1}$	0-2	2-4	4-6
n_i	20	30	50

6. Составить группированный статистический ряд по данным выборки:

30,2 27,5 18,3 31,4 10,9 27,5 20,1 40,4 29,3 14,6
32,1 36,7 29,3 11,6 27,6 22,9 29,3 28,4 31,1 21,9

7. Найти эмпирическую функцию по данному распределению выборки.

а)

x_i	2	5	7	8
n_i	1	3	2	4

б)

x_i	x_i	4	7	8
n_i	2	5	2	3

8. Группированная выборка распределения рейтинга успеваемости студентов (в баллах) представлена в таблице. Зачет получают студенты, набравшие более 300 баллов. Найти вероятность того, что студенты не получат зачет.

$x_i - x_{i+1}$	100-150	150-200	200-250	250-300	300-350	350-400	400-450
n_i	8	15	18	26	16	12	5

2. СТАТИСТИЧЕСКИЕ ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

2.1. Числовые характеристики генеральной и выборочной совокупностей

Пусть изучается дискретная генеральная совокупность относительно количественного признака *X*.

Генеральной средней называют среднее арифметическое значений признака генеральной совокупности.

$$\bar{x}_Г = \frac{x_1 N_1 + x_2 N_2 + \ldots + x_k N_k}{N} = \frac{1}{N}\sum_{i=1}^{k} x_i N_i$$

где x_i – *i*-я варианта; *Ni* – частота *i*-й варианты; *N* – объём генеральной совокупности.

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

Генеральной дисперсией называют среднее арифметическое квадратов отклонений значений признака X от генеральной средней.

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

Генеральное среднее квадратическое отклонение: $\sigma_Г = \sqrt{D_Г}$

Модой (Мо) называют варианту, которая имеет наибольшую частоту.

Медианой (m_e) называют варианту, которая делит вариационный ряд на две части, по числу вариант.

2.2. Статистические оценки

Пусть требуется изучить количественный признак генеральной совокупности. Допустим теоретически удалось установить, какое именно распределение имеет признак. Возникает задача оценки параметров, которыми определяется это распределение. Например, если известно, что исследуемый признак распределен нормально, то необходимо оценить (приближенно найти) математическое ожидание и среднеквадратическое отклонение, т.к. эти параметры полностью определяют нормальное распределение. Если есть основание считать, что признак имеет распределение Пуассона, то необходимо найти параметр λ.

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, ..., X_n, можно сказать, что найти статистическую оценку неизвестного параметра теоретического распределения – это значит найти функцию от наблюдаемых случайных величин, которая и дает приближенное значение оцениваемого параметра.

Статистической оценкой Θ^* неизвестного параметра Θ теоретического распределения называют функцию $f(x_1;x_2;\ldots x_n)$ от наблюдаемых СВ $X_1; X_2;\ldots, X_n$.

Пусть Θ^* – статистическая оценка неизвестного параметра Θ теоретического распределения. Допустим, что по выборке объема n найдена оценка Θ_1^*. Извлекаем из генеральной совокупности другую выборку того же объема и по ее данным находим оценку Θ_2^*. Получаем числа $\Theta_1^*, \Theta_2^*, \ldots, \Theta_k^*$, которые различны между собой. Таким образом, оценку Θ^* можно рассматривать как случайную величину, а числа $\Theta_1^*, \Theta_2^*, \ldots, \Theta_k^*$ – ее возможными значениями.

Точечной называют статистическую оценку, которая определяется одним числом $\Theta^* = f(x_1; x_2; \ldots x_n)$, где x_1, x_2, …, x_n – результаты наблюдений над количественным признаком (выборка).

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.

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

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

Однако несмещенность не является достаточным условием хорошего приближения к истинному значению оцениваемого параметра. Если при этом возможные значения Θ^* могут значительно отклоняться от среднего значения, то есть дисперсия Θ^* велика, то значение, найденное по данным одной выборки, может значительно отличаться от оцениваемого параметра. Следовательно, требуется наложить ограничения на дисперсию.

Состоятельной называется статистическая оценка, которая при $n \to \infty$ стремится по вероятности к оцениваемому параметру (если эта оценка несмещенная, то она будет состоятельной, если при $n \to \infty$ ее дисперсия стремится к 0).

2.3. Точечные оценки выборки

Вариационный ряд содержит достаточно полную информацию об изменчивости признака X. Однако на практике часто оказывается, что этого недостаточно и необходимо найти некоторые сводные характеристики ва-

риационных рядов: средних, центральной тенденции и изменчивости (показателей вариации), расчет которых представляет собой следующий этап после группировки и обработки данных наблюдений.

Для описания группирования и рассеивания наблюдаемых данных используются так называемые *числовые характеристики выборочной совокупности*. Эти числовые характеристики аппроксимируют соответствующие генеральные характеристики, т. е. являются их оценками. Таким образом, вместо числовых характеристик генеральной совокупности X достаточно рассмотреть аналогичные выборочные характеристики.

Пусть для изучения генеральной совокупности относительно некоторого количественно признака X произведена выборка объема n.

1) Выборочная средняя

Выборочной средней называют среднее арифметическое значение признака выборочной совокупности.

Если все значения признака выборки различны, то

$$\bar{x}_в = \frac{x_1 + x_2 + \ldots + x_k}{n}$$

если же все значения имеют частоты n_1, n_2, …, n_k, то

$$\bar{x}_в = \frac{x_1 n_1 + x_2 n_2 + \ldots + x_k n_k}{n} = \frac{1}{n}\sum_{i=1}^{k} x_i n_i$$

Выборочная средняя является несмещенной и состоятельной оценкой генеральной средней.

Замечание 1: Если выборка представлена интервальным вариационным рядом, то за x_i принимают середины частичных интервалов.

Замечание2. Если первоначальные варианты x_i – большие числа, то для упрощения расчета целесообразно вычесть из каждой варианты одно и то же число C, т. е. перейти к условным вариантам $u_i = x_i - C$ (в качестве C выгодно принять число, близкое к выборочной средней; поскольку выборочная средняя неизвестна, число C выбирают "на глаз"). Тогда

$$\bar{x}_в = C + \frac{1}{n}\sum_{i=1}^{k} x_i n_i = C + \bar{u}$$

2) Выборочная дисперсия

Для того, чтобы наблюдать рассеяние количественного признака значений выборки вокруг своего среднего значения, вводят сводную характеристику – выборочную дисперсию.

Выборочной дисперсией называют среднее арифметическое квадратов отклонения наблюдаемых значений признака от их среднего значения.

Если все значения признака выборки различны, то

$$D_в = \frac{1}{n}\sum_{i=1}^{k}\left(x_i - \overline{x}_в\right)$$

если же все значения имеют частоты n_1, n_2, ..., n_k, то

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

Так же, как в теории случайных величин, можно доказать, что справедлива следующая формула для вычисления выборочной дисперсии:

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

Для характеристики рассеивания значений признака выборки вокруг своего среднего значения пользуются сводной характеристикой – средним квадратическим отклонением.

Выборочным средним квадратическим отклонением называют квадратный корень из выборочной дисперсии: $\sigma_в = \sqrt{D_в}$

Величина характеризует $\sigma_в$ среднее значение отклонения вариант от выборочной средней без учета знака этого отклонения. Особенность состоит в том, что оно измеряется в тех же единицах, что и данные выборки.

Пример.

Найдем числовые характеристики выборки, заданной статистическим рядом

x_i	2	5	7	8
n_i	3	8	7	2

$$\overline{x}_в = \frac{1}{20}\left(2\cdot 3 + 5\cdot 8 + 7\cdot 7 + 8\cdot 2\right) = 5{,}5$$

$$D_в = \overline{x^2} - \left(\overline{x}\right)^2$$

$$\overline{x^2} = \frac{1}{20}\left(2^2\cdot 3 + 5^2\cdot 8 + 7^2\cdot 7 + 8^2\cdot 2\right) = 34{,}15$$

$$D_в = 34{,}15 - (5{,}55)^2 = 3{,}3475$$

$$\sigma_в = \sqrt{D_в} = \sqrt{3{,}3475} = 1{,}83$$

Замечание 1: если выборка представлена интервальным вариационным рядом, то за x_i принимают середины частичных интервалов.

Замечание 2. Если первоначальные варианты большие числа, то целесообразно вычесть из всех вариант одно и то же число C, равное выборочной средней или близкое к ней, т. е. перейти к условным вариантам $u_i = x_i - C$ (дисперсия при этом не изменится). Тогда $D_в = \overline{u^2} - \left(\overline{u}\right)^2$.

Замечание 3. Если первоначальные варианты являются десятичными дробями с k десятичными знаками после запятой, то, чтобы избежать действий с дробями, умножают первоначальные варианты на постоянное число $C = 10^k$, т. е. переходят к условным вариантам $u_i = Cx_i$. При этом дисперсия увеличится в C^2 раз. Поэтому, найдя дисперсию условных вариант, надо разделить ее на C^2:

$$D_в(X) = \frac{1}{n} D_в(u)$$

3) Исправленная дисперсия

Выборочная дисперсия является смещенной оценкой генеральной дисперсии, т. е. математическое ожидание выборочной дисперсии не равно оцениваемой генеральной дисперсии, а равно

$$M(D_в) = \frac{n-1}{n} D_Г,$$

где $D_Г$ – истинное значение дисперсии генеральной совокупности.

Для исправления выборочной дисперсии достаточно умножить ее на дробь $\frac{n}{n-1}$.

В качестве оценки генеральной дисперсии принимают **исправленную дисперсию s^2**, вычисляемую по формуле

$$s^2 = \frac{n}{n-1} D_в$$

Такая оценка будет являться несмещенной. Ей соответствует **исправленное среднее квадратическое отклонение** $s = \sqrt{s^2}$.

Замечание: формулы для вычисления выборочной дисперсии и исправленной дисперсии отличаются только знаменателями. При достаточно больших n выборочная и исправленная дисперсии мало отличаются. На практике исправленной дисперсией и исправленным средним квадратическим отклонением пользуются, когда объем выборки мал ($n < 30$). При большом объеме выборки ($n > 30$) параметры распределения оценивают по выборочным характеристикам.

$n < 30$	$n > 30$
$\bar{x}_{г} \approx \bar{x}_{в} \approx M(X)$	$\bar{x}_{г} \approx \bar{x}_{в} \approx M(X)$
$D(X) \approx s^2$	$D(X) \approx D_{в}$
$\sigma(X) \approx s$	$\sigma(X) \approx \sigma_{в}$

Пример.

Найти выборочную среднюю и выборочную дисперсию.

x_i	1	2	3	4
n_i	20	15	10	5

Решение.

$$\bar{x}_{в} = \frac{1 \cdot 20 + 2 \cdot 15 + 3 \cdot 10 + 4 \cdot 5}{50} = 2$$

$$D_{в} = \frac{20(1-2)^2 + 15(2-2)^2 + 10(3-2)^2 + 5(4-2)^2}{50} = 1$$

4) Коэффициент вариации

Коэффициент вариации применяют для сравнения вариации признаков сильно отличающихся по величине, или имеющих разные единицы измерения (разные наименования).

$$v = \frac{\sigma_{в}}{\bar{x}_{в}} \cdot 100\%$$

На практике считают, что если $v < 33\%$, то совокупность однородная.

Пример.

Предположим, что цены на ценные бумаги широко колеблются. Инвестор, который покупает акции по низкой цене, а продает по высокой, имеет хороший доход. Однако если цены на акции падают ниже стоимости, по которой инвестор купил, то он теряет доход.

За пять недель изменение цены составило:

на акции первого вида – \$ 57, 68, 64, 71, 62;

на акции 2 второго вида – \$ 12, 17, 8, 15, 13.

Чтобы оценить меру риска, инвестор может использовать коэффициент вариации и среднеквадратическое отклонение.

Какую информацию о степени риска может дать коэффициент вариации по сравнению со среднеквадратическим отклонением?

Решение.

Для акций первого вида: $\bar{x} = 64{,}4;\ s_x = 4{,}84$.

Для акций второго типа: $\bar{y} = 64{,}4;\ s_y = 3{,}03$.

Со среднеквадратическим отклонением как мерой риска акции первого типа более рискованные. Однако среднее арифметическое первых акций почти в 5 раз больше среднего арифметического вторых акций. Коэффициент вариации, используемый в данном случае, дает следующие результаты:

$$v_1 = \frac{4{,}84}{64{,}4} \cdot 100\% = 7{,}52\%$$

$$v_1 = \frac{3{,}03}{13} \cdot 100\% = 23{,}31\%$$

Для вторых акций коэффициент вариации почти в три раза больше, чем коэффициент вариации для первых акций. Таким образом, использование коэффициента вариации позволяет сделать заключение, что покупать акции второго типа более рискованно.

5) Мода

Модой M_0 называют варианту, которая имеет наибольшую частоту.

Пример.

Найти моду для данного статистического ряда.

варианта	1	4	7	9
частота	5	1	20	6

Решение.

Наиболее часто встречающаяся варианта имеет частоту 20, поэтому M_0= 7.

Для интервального вариационного сначала находят интервал группировки с наибольшей частотой (модальный интервал). Внутри модального интервала мода определяется по формуле:

$$M_0 = x_k + \frac{n_k - n_{k-1}}{2n_k - (n_{k-1} - n_{k+1})} h$$

где, x_k – нижняя граница модального интервала; n_k – частота указанного выше интервала; n_{k-1} – частота интервала, находящегося слева от модального интервала; n_{k+1} – частота интервала, находящегося справа от модального интервала.

6) Медиана.

Медианой вариационного ряда называется серединная точка в вариационном ряду, которая делит вариационный ряд на две равные по числу членов части. Для вариационного ряда медиана определяется в зависимости от того, является ли объем выборки *n* числом четным или нечетным.

Если число вариант нечетно, т. е. $n = 2l + 1$, то $m_e = x_{l+1}$. При четном $n = 2l$ медиана $m_e = \frac{x_{l.} + x_{l+1}}{2}$.

$$m_e = \begin{cases} x_{l+1}, \text{при } n = 2l + 1 \\ \frac{x_{l.+} x_{l+1}}{2}, \text{при } n = 2l \end{cases}$$

Например, для ряда: 2 3 5 6 7 m_e= 5; для ряда: 2 3 5 6 7 9 медиана равна m_e=(5+6)/2=5,5

Для интервального вариационного сначала находят интервал группировки, в котором содержится медиана, путем подсчета накопленных частот или накопленных относительных частот. Медианным будет тот интервал, в котором накопленная частота впервые окажется больше $n/2$. Внутри медианного интервала медиана определяется по следующей формуле:

$$m_e = x_k + \frac{0{,}5n - n_{k-1}}{n_{m_e}} h_{m_e}$$

Здесь x_k – нижняя граница медианного интервала; h_{m_e} – ширина медианного интервала; n_{k-1} – накопленная частота интервала, предшествующего медианному, n_{m_e} – частота медианного интервала.

Пример.

Для данного интервального статистического ряда определить моду и медиану.

Возрастные группы	До 20 лет	20–25	25–30	30–35	35–40	40–45	Старше 45
Число студентов	346	872	1054	781	212	121	76

Решение.

Объем выборки равен $n = 3462$.

Возрастные группы	Число студентов	Сумма накопленных частот
До 20 лет	346	346
20 - 25	872	1218
25 - 30	1054	2272
30 - 35	781	3053
35 -40	212	3265
40 - 45	121	3386
45 лет и более	76	3462
Итого	3462	

Модальный интервал находится в пределах возрастной группы 20–30 лет, так как в этом интервале находится наибольшая частота (1054).

Рассчитаем величину моды:

$$M_0 = 25 + \frac{1054 - 872}{2 \cdot 1054 - (872 - 781)} \cdot 5 = 27$$

Это значит, что модальный возраст студентов равен 27 годам.

Вычислим медиану. Медианный интервал находится в возрастной группе 25–30 лет, так как в пределах этого интервала расположена варианта, которая делит совокупность на две равные части:

$$m_e = 25 + \frac{0{,}5 \cdot 3462 - 1218}{1054} \cdot 5 = 27{,}5$$

Это означает, что половина студентов имеет возраст до 27,4 года, а другая – свыше 27,4 года.

Моду и медиану можно также определить графически.

Мода определяется по полигону или гистограмме (рис. 5) распределения. В первом случае мода соответствует наибольшей ординате. Во втором – правую вершину модального прямоугольника соединяют с правым углом предыдущего прямоугольника, а левую вершину – с левым углом последующего прямоугольника. Абсцисса точки пересечения – этих прямых будет модой распределения.

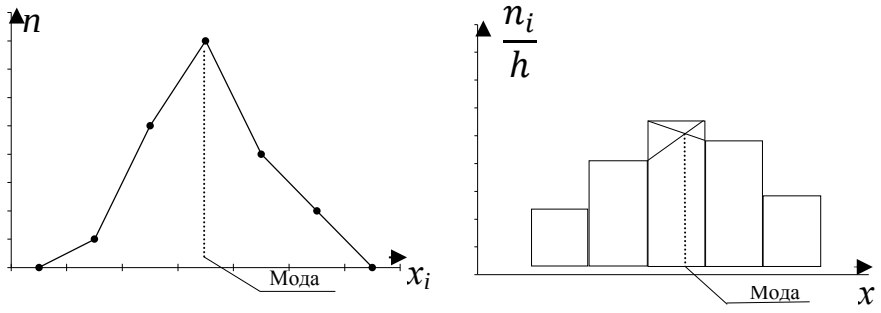


Рис. 5. Графическое определение моды

Медиана определяется по кумуляте (рис. 6). Для ее определения высоту наибольшей ординаты, которая соответствует общей численности совокупности, делят пополам. Через полученную точку проводят прямую, параллельную оси абсцисс, до пересечения ее с кумулятой. Абсцисса точки пересечения является медианой.

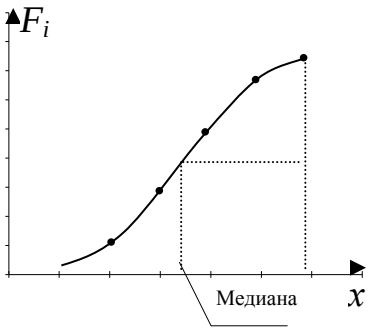


Рис. 6. Графическое определение медианы

2.4. Выборочные начальные и центральные моменты

Среднее выборочное и выборочная дисперсия являются частным случаем более общего понятия – *момента* статистического ряда.

Начальным выборочным моментом порядка l называется среднее арифметическое *l*-х степеней всех значений выборки:

$$v_l^* = \frac{1}{n}\sum_{i=1}^{m} x_i^l \cdot n_i$$

Из определения следует, что начальный выборочный момент первого порядка: $v_1^* = \frac{1}{n}\sum_{i=1}^{m} x_i \cdot n_i = \bar{x}_в$.

Центральным выборочным моментом порядка l называется среднее арифметическое *l*-х степеней отклонений наблюдаемых значений выборки от выборочного среднего $\bar{x}_в$:

$$\mu_l^* = \frac{1}{n}\sum_{i=1}^{m}\left(x_i - \bar{x}_в\right)^l \cdot n_i$$

Из определения следует, что *центральный выборочный момент второго порядка:*

$$\mu_2^* = \frac{1}{n}\sum_{i=1}^{m}\left(x_i - \bar{x}_в\right)^2 \cdot n_i = D_в = \sigma_в^2 .$$

Выборочным коэффициентом асимметрии называется число A_s^*, определяемое формулой: $A_s^* = \frac{\mu_3^*}{\sigma_в^3}$.

Выборочный коэффициент асимметрии служит для характеристики асимметрии полигона вариационного ряда. Если полигон асимметричен, то

одна из ветвей его, начиная с вершины, имеет более пологий "спуск", чем другая.

Если $A_s^* < 0$, то более пологий "спуск" полигона наблюдается слева; если $A_s^* > 0$ – справа. В первом случае асимметрию называют *левосторонней*, а во втором – *правосторонней*.

Выборочным коэффициентом эксцесса или *коэффициентом крутости* называется число E_k^*, определяемое формулой:

$$E_k^* = \frac{\mu_4^*}{\sigma_в^4} - 3 .$$

Выборочный коэффициент эксцесса служит для сравнения на "крутость" выборочного распределения с нормальным распределением.

Коэффициент эксцесса для случайной величины, распределенной по нормальному закону, равен нулю.

Поэтому за стандартное значение выборочного коэффициента эксцесса принимают $E_k^* = 0$.

Если $E_k^* < 0$, то полигон имеет более пологую вершину по сравнению с нормальной кривой; если $E_k^* > 0$, то полигон более крутой по сравнению с нормальной кривой.

При вычислении числовых характеристик выборки можно использовать таблицу

x_i	n_i	$x_i \cdot n_i$	$x_i - \bar{x}_в$	$(x_i - \bar{x}_в)^2 \cdot n_i$	$(x_i - \bar{x}_в)^3 \cdot n_i$	$(x_i - \bar{x}_в)^4 \cdot n_i$
x_1	n_1	$x_1 \cdot n_1$	$x_1 - \bar{x}_в$	$(x_1 - \bar{x}_в)^2 \cdot n_1$	$(x_1 - \bar{x}_в)^3 \cdot n_1$	$(x_1 - \bar{x}_в)^4 \cdot n_1$
…	…	…	…	…	…	…
x_m	n_m	$x_m \cdot n_m$	$x_m - \bar{x}_в$	$(x_m - \bar{x}_в)^2 \cdot n_m$	$(x_m - \bar{x}_в)^3 \cdot n_m$	$(x_m - \bar{x}_в)^4 \cdot n_m$
$\sum$						

С помощью суммы $\sum_{i=1}^{m} x_i \cdot n_i$ находим $\bar{x}_в$;

с помощью суммы $\sum_{i=1}^{m} (x_i - \bar{x}_в)^2 \cdot n_i$ находим $D_в$ и $\sigma_в$;

с помощью суммы $\sum_{i=1}^{m} (x_i - \bar{x}_в)^3 \cdot n_i$ находим A_s^*;

с помощью суммы $\sum_{i=1}^{m} (x_i - \bar{x}_в)^4 \cdot n_i$ находим E_k^*.

2.5. Метод условных вариант для расчета характеристик выборки

Предположим, что варианты выборки расположены в возрастающем порядке, т. е. в виде вариационного ряда. Если численные значения вариант и их частот велики, то вычисления числовых характеристик могут быть громоздкими. В этом случае, как было сказано ранее, переходят к условным вариантам.

Для рядов с равноотстоящими вариантами используют метод расчета сводных характеристик, основанный на замене выборочных вариант другими небольшими числами условными вариантами – u_i.

Формула перехода к условным вариантам: $u_i = \dfrac{x_i - c}{h}$, где c – новое начало отсчёта (ложный нуль); h – шаг таблицы, разность между двумя соседними вариантами (длина частичного интервала). Числа c и h выбираются произвольно. Чтобы упростить вычисления в качестве c выбирают вариант, который имеет наибольшую частоту или находится в середине ряда.

Вариационный ряд признака X заменяется вариационным рядом признака U. Для последнего находят числовые характеристики $\bar{u}_в$, D_u, σ_u, а затем переходят к первоначальным вариантам по формулам перехода:

а*) выборочная средняя:* $\bar{x}_в = \bar{u}_в \cdot h + c$

б) *выборочная дисперсия*: $D_в = D_u \cdot h^2$

в) среднее квадратическое отклонение: $\sigma_в = \sigma_u \cdot h = \sqrt{D_в}$

*Пример***.**

Для приведенного распределения вычислить основные выборочные характеристики.

x_i	0	1	2	3	4	5	6	7
n_i	5	12	18	25	14	5	4	2

Решение.

Для вычисления основных числовых характеристик воспользуемся упрощающими формулами. Выберем c = 3; h = 1. Составим расчетную таблицу.

x_i	n_i	u_i	$u_i \cdot n_i$	$u_i^2 \cdot n_i$
0	5	–3	–15	45
1	12	–2	–24	48
2	18	–1	–18	18
3	25	0	0	0
4	14	1	14	14
5	5	2	10	20
6	4	3	12	36
7	2	4	8	32
Сумма	**80**	**–**	**–13**	**213**

$$\bar{u}_в = \frac{1}{n}\sum_{i=1}^{m} u_i \cdot n_i = \frac{-13}{80} = -0{,}1625$$

$$D_u = \frac{1}{n}\sum_{i=1}^{m} u_i^2 \cdot n_i - \left(\bar{u}_в\right)^2 = \frac{213}{80} - \left(-0{,}1625\right)^2 \approx 2{,}64$$

$$\bar{x}_в = \bar{u}_в \cdot h + c = -0{,}1625 + 3 = 2{,}8375 \approx 2{,}84$$

$$D_в = D_u \cdot h^2 = 2{,}64 \cdot 1^2 \approx 2{,}64$$

$$\sigma_в = \sqrt{D_в} \approx 1{,}62$$

2.6. Интервальные оценки

При выборке малого объема точечная оценка может значительно отличаться от оцениваемого параметра, что приводит к грубым ошибкам. Поэтому в таком случае лучше пользоваться ***интервальными оценками,*** то есть указывать интервал, в который с заданной вероятностью попадает истинное значение оцениваемого параметра. Разумеется, чем меньше длина этого интервала, тем точнее оценка параметра. Поэтому, если для оценки Θ^* некоторого параметра Θ справедливо неравенство $\left|\Theta^* - \Theta\right| < \delta$, число $\delta > 0$ характеризует **точность оценки** (чем меньше δ, тем точнее оценка). Но статистические методы позволяют говорить только о том, что это неравенство выполняется с некоторой вероятностью.

Надежностью (доверительной вероятностью) оценки Θ^* параметра Θ называется вероятность γ того, что выполняется неравенство $\left|\Theta^* - \Theta\right| < \delta$.

Если заменить это неравенство двойным неравенством

$$-\delta < \Theta^* - \Theta < \delta,$$

то получим:

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

Таким образом, γ есть вероятность того, что Θ попадает в интервал $\left(\Theta^* - \delta; \Theta^* + \delta\right)$.

Доверительным называется интервал, в который попадает неизвестный параметр с заданной надежностью γ.

2.7. Построение доверительных интервалов

1. Интервальная оценка (с надежностью γ) математического ожидания нормального распределения при известной дисперсии.

Пусть исследуемая случайная величина X распределена по нормальному закону с известным средним квадратическим σ, и требуется по значению выборочного среднего $\bar{x}_в$ оценить ее математическое ожидание a. Будем рассматривать выборочное среднее $\bar{x}_в$ как случайную величину $\overline{X}$, а значения вариант выборки $x_1, x_2,\ldots, x_n$ как одинаково распределенные независимые случайные величины $X_1, X_2,\ldots, X_n$, каждая из которых имеет математическое ожидание a и среднее квадратическое отклонение σ. При этом $M(\overline{X}) = a$, $\sigma(\overline{X}) = \dfrac{\sigma}{\sqrt{n}}$ (используем свойства математического ожидания и дисперсии суммы независимых случайных величин). Оценим вероятность выполнения неравенства $|\overline{X} - a| < \delta$. Применим формулу для вероятности попадания нормально распределенной случайной величины в заданный интервал:

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

Тогда, с учетом того, что $\sigma(\overline{X}) = \dfrac{\sigma}{\sqrt{n}}$,

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

Отсюда $\delta = \dfrac{t\sigma}{\sqrt{n}}$, и предыдущее равенство можно переписать так:

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

Итак, значение математического ожидания a с вероятностью (надежностью) γ попадает в интервал $\left(\bar{x}_в - \dfrac{t\sigma}{\sqrt{n}} < a < \bar{x}_в + \dfrac{t\sigma}{\sqrt{n}}\right)$, где значение t определяется из таблиц для функции Лапласа так, чтобы выполнялось равенство $2\Phi(t) = \gamma$.

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

Пример.

Найдем доверительный интервал для математического ожидания нормально распределенной случайной величины, если объем выборки $n = 49$, $\bar{x}_в = 2{,}8$, $\sigma = 1{,}4$; а доверительная вероятность $\gamma = 0{,}9$.

Решение.

По условию: $n = 49$; $\bar{x}_в = 2{,}8$; $\sigma = 1{,}4$; $\gamma = 0{,}9$.

По таблице значений интегральной функции Лапласа (Приложение 2) определим t, при котором $2\Phi(t) = \gamma$.

$2\Phi(t) = 0{,}9 \Rightarrow \Phi(t) = 0{,}45 \Rightarrow t = 1{,}645$, тогда

$$2{,}8 - \frac{1{,}645 \cdot 1{,}4}{\sqrt{49}} < a < 2{,}8 + \frac{1{,}645 \cdot 1{,}4}{\sqrt{49}}$$

или $2{,}471 < a < 3{,}129$.

Найден доверительный интервал, в который попадает a с надежностью 0,9.

2. Интервальная оценка (с надежностью γ) математического ожидания нормального распределения при неизвестной дисперсии.

Если известно, что исследуемая случайная величина X распределена по нормальному закону с неизвестным средним квадратическим отклонением, то

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

Получаем доверительный интервал для математического ожидания a:

$$\overline{x_в} - t_\gamma \frac{s}{\sqrt{n}} < a < \overline{x_в} + t_\gamma \frac{s}{\sqrt{n}},$$

где s – исправленное выборочное квадратическое отклонение, t_γ можно найти по соответствующей таблице при заданных n и γ.

Пример.

Пусть объем выборки $n = 25$, $\bar{x}_в = 3$; $s = 1{,}5$. Найти доверительный интервал для математического ожидания с надежностью γ=0,99.

Решение.

Найдем доверительный интервал для a при $\gamma = 0{,}99$.

Из таблицы (Приложение 3) находим, что $t_\gamma (n = 25, \gamma = 0{,}99) = 2{,}797$.

Тогда

$$3 - 2{,}797 \cdot \frac{1{,}5}{\sqrt{25}} < a < 3 + 2{,}797 \cdot \frac{1{,}5}{\sqrt{25}}$$

или $2{,}161 < a < 3{,}839$ – доверительный интервал, в который попадает a с вероятностью 0,99.

3. Интервальная оценка (с надежностью γ) среднего квадратического отклонения нормального распределения.

Будем искать для среднего квадратического отклонения нормально распределенной случайной величины доверительный интервал вида $(s-\delta; s+\delta)$, где s – исправленное выборочное среднее квадратическое отклонение, а для δ выполняется условие: $P(|\sigma - s| < \delta) = \gamma$.

Предположим, что, тогда

$s(1-q) < \sigma < (1+q)$, если $q < 1$

$0 < \sigma < (1+q)$, если $q > 1$

q находят по таблице (Приложение 4) по заданным n и γ.

Пример.

Пусть $n = 20$, $s = 1{,}3$. Найдем доверительный интервал для σ при заданной надежности $\gamma = 0{,}95$.

Из соответствующей таблицы находим q ($n = 20$; $\gamma = 0{,}95$) = 0,37.

Следовательно, границы доверительного интервала:

$1{,}3(1-0{,}37) < \sigma < 1{,}3(1+0{,}37)$

Итак, $0{,}819 < \sigma < 1{,}781$ с вероятностью 0,95.

Вопросы для самоконтроля

1. Дайте определение точечной статистической оценки.
2. Какими свойствами обладает выборочное среднее?
3. Какими свойствами обладает выборочная дисперсия?
4. Дайте определение моды и медианы выборки.
5. Какая оценка параметра распределения называется точечной?
6. Какая числовая характеристика выборки является несмещенной для математического ожидания?
7. Какая числовая характеристика выборки является несмещенной для дисперсии?
8. Что понимается под термином "интервальная оценка параметра распределения"?
9. Дайте определение доверительного интервала.
10. Что называется доверительной вероятностью? Какие значения она принимает?
11. Что такое точность оценки и надежность оценки?
12. Как изменится длина доверительного интервала, если увеличить: 1) объем выборки, 2) доверительную вероятность? Ответ обоснуйте.

13. Запишите формулу для нахождения доверительного интервала математического ожидания нормально распределенной случайной величины. если генеральная дисперсия: 1) известна: 2) неизвестна.

Задачи для самостоятельного решения

1. Из генеральной совокупности извлечена выборка объема $n = 50$. Найти несмещенную оценку генеральной средней.

x_i	2	5	7	10
n_i	16	12	8	14

2. Найти выборочную среднюю по данному распределению выборки объема $n = 10$:

x_i	1250	1270	1280
n_i	2	5	3

3. Найти выборочную дисперсию для выборки.

а)

x_i	186	192	194
n_i	2	5	3

б)

x_i	0,01	0,04	0,08
n_i	5	3	2

4. Найти выборочную дисперсию для выборки.

x_i	1250	1275	1280	1300
n_i	20	25	50	5

5. Ниже приведены результаты измерения роста (в см) случайно отобранных 100 студентов. Найти выборочную среднюю и выборочную дисперсию роста студентов.

Рост	154-158	158-162	162-166	166-170	170-174	174-178	178-182
Число студентов	10	14	26	28	12	8	2

6. Дан вариационный ряд выборки объема $n = 10$: –2, 0, 3, 3, 4, 5, 9, 11, 12, 15. Найти медиану и моду для этого ряда.

7. Найти медиану выборки, заданной таблицей:

Интервал	-1 – 0	0 – 1	1 – 2	2 – 3
Частота	30	70	80	20

8. Найти медиану и моду выборки, заданной таблицей:

Рост	154-158	158-162	162-166	166-170	170-174	174-178	178-182
Число студентов	10	14	26	28	12	8	2

9. Путем опроса получены следующие данные (n = 80):

2 4 2 4 3 3 3 2 0 6 1 2 3 2 2 4 3 3 5 1 0 2 4 3 2 2 3 3 1 3 3 3 1 1 2 3 1 4 3 1 7
4 3 4 2 3 2 3 3 1 4 3 1 4 5 3 4 2 4 5 3 6 4 1 3 2 4 1 3 1 0 0 4 6 4 7 4 1 3 5 1.

Требуется:

1) Составить статистическое распределение выборки, предварительно записав дискретный вариационный ряд.

2) Найти основные числовые характеристики вариационного ряда: выборочное среднее; выборочную дисперсию; выборочное среднее квадратическое отклонение; коэффициент вариации; моду; медиану.

3) Пояснить смысл полученных результатов.

10. Найти доверительный интервал для оценки с надежностью 0,99 неизвестного математического ожидания a нормально распределенного признака X генеральной совокупности, если известны генеральное ско $\sigma = 5$, выборочная средняя $\bar{x}_в = 10{,}2$ и объем выборки $n = 16$.

11. Найти минимальный объем выборки, при котором с надежностью 0,975 точность оценки математического ожидания a генеральной совокупности по выборочной средней равна $\delta = 0{,}3$, если известно среднее квадратическое отклонение $\sigma = 1{,}2$ нормально распределенной генеральной совокупности.

12. Из генеральной совокупности извлечена выборка объема n = 12:

x_i	-0,5	-0,4	-0,2	0	0,32	0,6	0,8	1	1,2	1,5
n_i	1	2	1	1	1	1	1	1	2	1

Оценить с надежностью 0,95 математическое ожидание a нормально распределенного признака генеральной совокупности с помощью доверительного интервала.

13. Произведено 12 измерений одним прибором (без систематической ошибки) некоторой физической величины, причем "исправленное" среднее квадратнческое отклонение s случайных ошибок измерений оказалось равным 0,6. Найти точность прибора с надежностью 0,99. Предполагается, что результаты измерений распределены нормально.

14. С целью определения среднего трудового стажа на предприятии методом случайной повторной выборки проведено обследование трудового стажа рабочих. Из всего коллектива рабочих завода случайным образом выбрано 400 рабочих, данные о трудовом стаже которых и составили выборку. Средний по выборке стаж оказался равным 9,4 года. Считая, что трудовой стаж рабочих имеет нормальный закон распределения, определить с вероятностью 0,97 границы, в которых окажется средний трудовой стаж для всего коллектива, если известно, что σ = 1,7 года.

15. С целью определения средней продолжительности рабочего дня на предприятии методом случайной повторной выборки проведено обследование продолжительности рабочего дня сотрудников. Из всего коллектива завода случайным образом выбрано 30 сотрудников. Данные табельного учета о продолжительности рабочего дня этих сотрудников и составили выборку. Средняя по выборке продолжительность рабочего дня оказалась равной 6,85 часа, а s = 0,7 часа. Считая, что продолжительность рабочего дня имеет нормальный закон распределения, с надежностью γ = 0,95 определить, в каких пределах находится действительная средняя продолжительность рабочего дня для всего коллектива данного предприятия.

16. Результаты исследования длительности оборота оборотных средств торговых фирм города (в днях) представлены в группированном виде. Построить доверительный интервал с надежностью 0,95 для средней длительности оборота оборотных средств торговых фирм города при условии, что среднеквадратическое отклонение известно и равно 10 дням.

Интервал	14-23	23-32	32-41	41-50	50-59	59-68	68-77
Частота	2	3	9	17	10	6	3

17. Ниже приведены объемы выработки за месяц (в тыс. руб.) пятидесяти продавцов молочных изделий, работающих в разных районах города.

15	19	6	18	21	16	20	17	15	10
16	20	7	19	22	17	21	19	16	11
19	10	8	18	20	8	18	16	20	12
16	21	21	9	19	19	14	18	19	19
12	20	20	8	13	10	18	17	22	18

Требуется найти:

1) выборочную среднюю

2) исправленное среднеквадратическое отклонение

3) с надежность. γ = 0,95 – доверительный интервал для математического ожидания.

4) Построить гистограмму и эмпирическую функцию распределения.

18. Обследуется 25 растений пшеницы по числу зерен содержащихся в каждом колосе. Для удобства каждому растению присвоен номер. Числа зерен показаны в таблице.

27	28	28	24	25	27	29	29	27	28	30	27	27
33	28	30	28	29	28	29	25	29	29	28	28	

По имеющимся выборочным данным изучаемого признака (изучаемой случайной величины *X*) *выполнить следующие действия*.

1) Составить вариационный ряд.

2) Определить эмпирическую функцию распределения. Построить ее график.

3) Построить полигоны частот или относительных частот. Сделать вывод о законе распределения изучаемой величины.

4) Найти числовые характеристики изучаемой величины.

5) Найти моду, медиану выборки.

6) Найти точечные оценки математического ожидания, дисперсии, среднего квадратического отклонения изучаемой случайной величины.

7) Найти интервальную оценку математического ожидания с доверительной вероятностью 0,95.

3. КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

3.1. Основные понятия

Одна из задач статистического исследования состоит в изучении взаимозависимости между наблюдаемыми явлениями. Знание взаимозависимостей случайных величин дает возможность решить одну из кардинальных задач любого исследования: возможность предвидеть, прогнозировать развитие ситуации при изменении конкретных характеристик объекта исследования. Признаки по их сущности и значению для изучения взаимосвязи делятся на два класса. Признаки, обуславливающие изменения других, связанных с ними признаков, называются **факторными**, или просто факторами. Признаки, изменяющиеся под действием факторных признаков, называются **результативными**. Различают два типа взаимосвязей между различными явлениями и их признаками: *функциональная зависимость и статистическая зависимость* (либо независимость).

Если каждому определенному значению факторного *X* признака соответствует по определенному закону вполне определенное значение результативного признака *Y*, то такой вид причинной зависимости называется *функциональной связью*. В жизни функциональные связи встречаются далеко не всегда, поскольку каждая переменная определяется воздействием многих факторов – генетических, социальных, педагогических и т. д.

Известно, например, что между ростом и массой человека существует положительная связь: более высокие индивиды имеют обычно и большую массу, чем индивиды низкого роста. То же наблюдается и в отношении качественных признаков: блондины, как правило, имеют голубые, а брюнеты – карие глаза. Однако из этого правила имеются исключения, когда сравнительно низкорослые индивиды оказываются тяжелее высокорослых, и среди населения хотя и нечасто, но встречаются кареглазые блондины и голубоглазые брюнеты. Причина таких "исключений" в том, что каждый биологический признак, выражаясь математическим языком, является функцией многих переменных; на его величине сказывается влияние и генетических и средовых факторов, в том числе и случайных, что вызывает варьирование признаков. Отсюда зависимость между ними приобретает не функциональный, а ***статистический характер***, когда определенному значению одного признака, рассматриваемого в качестве независимой переменной, соответствует не одно и то же числовое значение, а целая гамма распределяемых в вариационный ряд числовых значений другого признака, рассматриваемого в качестве независимой переменной.

Статистической (стохастической) называется зависимость, при которой изменение одной случайной величины влечет изменение *распределения* другой случайной величины.

Частным случаем стохастической связи является **корреляционная** связь, при которой изменение среднего значения результативного признака обусловлено изменением факторных признаков. Термин "корреляция" происходит от лат. correlatio – соотношение, связь.

Например, дети, которые чаще смотрят по телевизору боевики, меньше читают. Дети, которые больше читают, лучше учатся. Не так-то просто решить, где тут причины, а где следствия, но это и не является задачей статистики. Статистика может лишь, выдвинув гипотезу о наличии связи, подкрепить ее цифрами. При этом *данный вид взаимосвязи между признаками проявляется в том, что при изменении одной из величин изменяется среднее значение другой.* Если увеличение одной случайной величины связано с увеличением второй случайной величины, корреляция называется прямой. Например, количество прочитанных страниц за год и средний балл (успеваемость). Если, напротив рост одной величины связано с уменьшением другой, говорят об обратной корреляции. Например, количество боевиков и количество прочитанных страниц.

Если функциональные связи одинаково легко обнаружить и на единичных, и на групповых объектах, то этого нельзя сказать о связях корреляционных, которые изучаются только на групповых объектах методами математической статистики.

Корреляционный анализ – это группа статистических методов, направленная на выявление и математическое представление структурных зависимостей между выборками.

Корреляция рассматривается как признак, указывающий на взаимосвязь ряда числовых последовательностей случайных величин. Иначе говоря, корреляция характеризует силу взаимосвязи в данных. Корреляционный анализ состоит в определении степени тесноты корреляционной связи между переменными и количественной оценке тесноты этой связи. Корреляционный анализ следует применять только в том случае, если данные наблюдений или эксперимента можно считать *случайными* и выбранными из *нормальной* совокупности.

Особенности корреляционной связи.

– Корреляционная связь не может рассматриваться как свидетельство причинно-следственной зависимости. Она свидетельствует лишь о том, что изменения одного признака, как правило, сопровождаются определенными изменениями другого, т. е. отражает согласованные изменения признаков, которые могут объясняться множеством причин, в том числе зависимостью обоих признаков от третьего признака или сочетания других признаков.

– Корреляционная связь не дает ответа на вопрос, где находится причина изменений – в одном из признаков или за пределами исследуемой пары признаков.

Задача корреляционного анализа сводится к

1) установлению направления и формы связи между признаками,
2) измерению ее тесноты
3) оценке достоверности выборочных показателей корреляции.

Корреляционные связи различаются **по форме, направлению и степени (силе).**

По форме корреляционная связь может быть прямолинейной или криволинейной.

По направлению корреляционная связь может быть положительной ("прямой") и отрицательной ("обратной"). При положительной прямоли-

нейной корреляции более высоким значениям одного признака соответствуют более высокие значения другого, а более низким значениям одного признака – низкие значения другого. При отрицательной корреляции соотношения обратные. При положительной корреляции коэффициент корреляции имеет положительный знак, при отрицательной корреляции – отрицательный знак.

Степень, сила или теснота корреляционной связи определяется по величине коэффициента корреляции. Сила связи не зависит от ее направленности и определяется по абсолютному значению коэффициента корреляции. Коэффициент корреляции, определяемый по выборочным данным, называется *выборочным коэффициентом.* Данный коэффициент впервые использовал Карл Пирсон (1857–1936), английский математик, разработавший статистический аппарат для проверки теории Ч. Дарвина. Термин "корреляция" был введен в науку выдающимся английским естествоиспытателем Ф. Гальтоном в 1886 г. Однако точную формулу для подсчёта коэффициента корреляции разработал его ученик К. Пирсон.

Коэффициент характеризует наличие только линейной связи между признаками, обозначаемыми, как правило, символами *X* и *Y*. Формула расчёта коэффициента корреляции построена таким образом, что, если связь между признаками имеет линейный характер, коэффициент Пирсона точно устанавливает тесноту этой связи. Поэтому он называется также коэффициентом линейной корреляции Пирсона. Если же связь между переменными *X* и *Y* не линейна, то Пирсон предложил для оценки тесноты этой связи так называемое корреляционное отношение.

Задача выявления связи между факторным и результативным признаками может быть решена при помощи следующих приёмов:

– визуализации связи (построение и визуальный анализ корреляционного поля);

– использования результатов аналитической группировки и др.

При использовании результатов аналитической группировки связь считается установленной, если группировка показывает изменение среднего значения результативного признака в группах при изменении факторного признака (основания группировки).

Графически взаимосвязь двух признаков изображается с помощью поля корреляции (диаграммы рассеяния). В системе координат на оси абсцисс откладываются значения факторного признака, а на оси ординат – ре-

зультативного. Каждое пересечение линий, проводимых через эти оси, обозначаются точкой. При отсутствии тесных связей имеет место беспорядочное расположение точек на графике. Чем сильнее связь между признаками, тем теснее будут группироваться точки вокруг определенной линии, выражающей форму связи.

Регрессия тесно связана с корреляцией и позволяет исследовать аналитическое выражение взаимосвязи между признаками.

Понятие "регрессия" связано с Ф. Гальтоном. В 1885 году был издан его научный труд "Регрессия в направлении к общему среднему размеру при наследовании роста". В этой работе он пришел к выводу, что признаки родителей не полностью наследуются детьми, и чем отдаленнее предок, тем в меньшей мере сказываются его свойства на потомке. Гальтон показал, что дети очень высоких или очень низких родителей в среднем имеют менее высокий или соответственно менее низкий рост. Кроме того, отклонение роста детей не так велико, как отклонение роста их родителей от среднего роста исследованных лиц. Это движение назад в направлении к среднему Гальтон назвал регрессией (*to regress* – движение в обратном направлении). Гальтон писал: "Закон регрессии веско свидетельствует против полного наследования какого-либо признака. Из большого числа детей только немногие будут уклоняться от среднего уровня по сравнению с уклонением одного из родителей, отличающегося своими природными качествами. Чем ярче талант одного из родителей, тем реже родители имеют счастье видеть, что природа также щедро одарила их сыновей, и еще реже бывает, чтобы одаренность передавалась в последующие поколения. Закон беспристрастен и объективен. Он равномерно распределяет наследование хороших и плохих признаков. Он разрушает чрезмерные иллюзии одного одаренного родителя, лелеющего мечту, что его дети унаследуют все его способности. Закон устраняет также преувеличенные опасения относительно того, что детям передадуться все слабости, недостатки и болезни родителей. Разумеется, эти утверждения не находятся в противоречии с общей теорией, согласно которой дети талантливых родителей имеют бóльшую вероятность обладать какими-либо дарованиями, чем дети родителей со средними способностями. Наши рассуждения выражают только тот факт, что самый одаренный из всех детей немногих высокоодаренных родительских пар не так будет талантлив, как самый одаренный из всех детей очень многих родительских пар со средними способностями."

В статистической трактовке регрессией называют изменение функции в зависимости от изменений одного или нескольких аргументов. Под функцией понимают переменную, которая зависит от другой переменной – аргумента (независимая переменная). Регрессия – это односторонняя статистическая зависимость. При простой корреляции изучают зависимость между изменчивостью двух переменных X и Y. С помощью регрессии ставится дополнительная задача: установить, как количественно меняется одна переменная при изменении другой (или других) на единицу. Если исследуют зависимость переменной Y от X, то устанавливают регрессию Y на X. Если же изучают зависимость переменной X от Y, то определяют регрессию X на Y. Цель регрессионного анализа – по значениям одной переменной, выбранной в качестве аргумента, предсказать соответствующее значение другой (функции). В этом заключается первое отличие метода регрессии от метода корреляции. Второе отличие состоит в том, что степень и характер регрессии можно установить и при небольшом числе пар значений зависимой и независимой переменных.

Регрессионный анализ заключается в определении аналитического выражения связи, в котором изменение одной величины (называемой зависимой или результативным признаком), обусловлено влиянием одной или нескольких независимых величин (факторных признаков).

Виды регрессии

1) Относительно числа учитываемых признаков регрессия может быть: между зависимой переменной Y и несколькими независимыми (объясняющими) переменными: X_1, X_2, …X_m.

2) Относительно формы зависимости регрессия может быть линейной и нелинейной.

3) Относительно направления связи: положительной отрицательной.

4) По характеру отношений между зависимой и независимыми переменными регрессия может быть непосредственной (причина оказывает прямое воздействие на следствие), косвенной (независимая переменная действует через какую-то третью или ряд других причин на зависимую переменную), ложной (нонсенс-регрессия – возникает при формальном подходе без уяснения причин, которые обуславливают данную связь).

Одной из проблем построения уравнений регрессии является их размерность, то есть определение числа факторных признаков, включаемых

в модель. Их число должно быть оптимальным. Сокращение размерности за счет исключения второстепенных, несущественных факторов позволяет получить модель, быстрее и качественнее реализуемую.

В то же время, построение модели малой размерности может привести к тому, что она будет недостаточно полно описывать исследуемое явление или процесс.

При построении моделей регрессии должны соблюдаться следующие требования:

1. Совокупность исследуемых исходных данных должна быть однородной и математически описываться непрерывными функциями.

2. Возможность описания моделируемого явления одним или несколькими уравнениями причинно-следственных связей.

3. Все факторные признаки должны иметь количественное (числовое) выражение.

4. Наличие достаточно большого объема исследуемой совокупности (в последующих примерах в целях упрощения изложения материала – это условие нарушено, т. е. объем очень мал).

5. Причинно-следственные связи между явлениями и процессами должны описываться линейной или приводимой к линейной форме зависимостью.

6. Отсутствие количественных ограничений на параметры модели связи.

7. Постоянство территориальной и временной структуры изучаемой совокупности.

Соблюдение данных требований позволяет построить модель, наилучшим образом описывающую реальные социально-экономические явления и процессы.

Корреляционное поле

Наличие качественной корреляционной связи между двумя исследуемыми числовыми наборами экспериментальных данных, можно обнаружить, изображая *поля корреляции.*

Когда исследуется корреляция между количественными признаками, значения которых можно точно измерить в единицах метрических шкал (метры, секунды, килограммы и т. д.), то очень часто принимается модель двумерной нормально распределенной генеральной совокупности. Такая

модель отображает зависимость между переменными величинами x_i и y_i графически в виде геометрического места точек в системе прямоугольных координат. Эту графическую зависимость называются также *диаграммой рассеивания* или ***корреляционным полем***.

Корреляционное поле и *корреляционная таблица* являются исходными данными при корреляционном анализе. Пусть $(x_k; y_k), k = 1,2,\ldots,n$ – результаты парных наблюдений над случайными величинами *X* и *Y*. Изображая полученные результаты в виде точек в декартовой системе координат, получим корреляционное поле. По характеру расположения точек поля можно составить предварительное представление о форме зависимости случайных величин (например, о том, что одна из них в среднем возрастает или убывает с возрастанием другой).

Данная модель двумерного нормального распределения (корреляционное поле) позволяет дать наглядную графическую интерпретацию коэффициента корреляции, т.к. распределение в совокупности зависит от пяти параметров: m_x, m_y – средние значения (математические ожидания); s_x,s_y – стандартные отклонения случайных величин *X* и *Y* и r_{xy} – коэффициент корреляции, который является мерой связи между случайными величинами *X* и *Y*.

Визуальный анализ корреляционного поля помогает выявить не только наличия статистической зависимости (линейную или нелинейную) между исследуемыми признаками, но и ее тесноту и форму. Это имеет существенное значение для следующего шага в анализе s выбора и вычисления соответствующего коэффициента корреляции.

Так, коэффициент линейной корреляции – это мера "плотности расположения" точек по отношению к некоторой прямой (т. е. линии). Коэффициент корреляции ***тем выше,*** чем ***с большей плотностью*** точки сосредоточены относительно ***наклонной*** прямой – как бы "вырисовывают" её, обнаруживая при этом линейную зависимость между двумя величинами.

1) Если $r_{xy} = 0$, то значения, x_i, y_i, полученные из двумерной нормальной совокупности, расположены на графике крайне хаотично, располагаются в координатах *x, y* в пределах области, ограниченной окружностью (рисунок 7). Нельзя выделить никакой закономерности между величинами *X* и *Y*. В этом случае между случайными величинами *X* и *Y* отсутствует корреляция и они называются некоррелированными. Для двумерного нор-

мального распределения некоррелированность означает одновременно и независимость случайных величин *X* и *Y*.

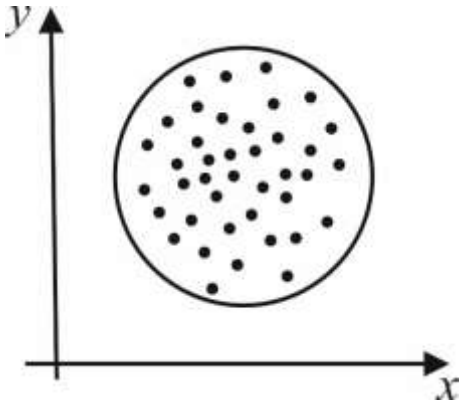
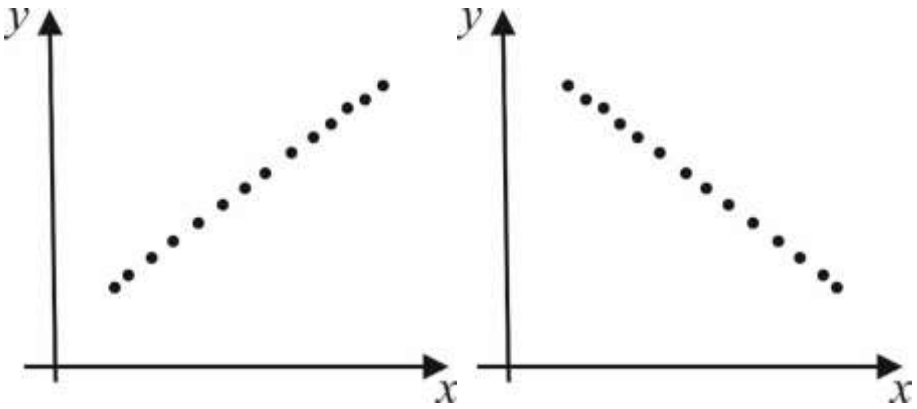


Рис. 7. Отсутствие корреляции между случайными величинами *X* и *Y*

2) Если $r_{xy} = \pm 1$, то между случайными величинами *X* и *Y* существует линейная функциональная зависимость. В этом случае говорят о полной корреляции. При $r_{xy} = 1$ (см. рисунок 8а) значения x_i, y_i определяют точки, лежащие на прямой линии, имеющей положительный наклон (с увеличением x_i значения y_i также увеличиваются), при $r_{xy} = -1$ (см. рис. 8б) прямая имеет отрицательный наклон.

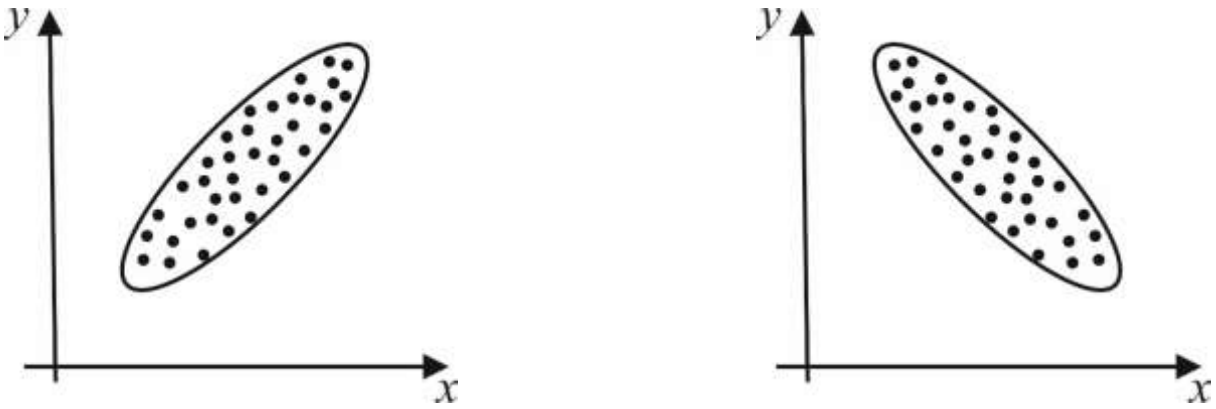


а) б)

Рис. 8. Линейная функциональная зависимость

3) В промежуточных случаях $(-1 < r_{xy} < 1)$ точки, соответствующие значениям x_i,y_i, попадают в область, ограниченную некоторым эллипсом, причем при $r_{xy} > 0$ имеет место положительная корреляция (с увеличением x_i значения y_i имеют тенденцию к возрастанию), при $r_{xy} < 0$ корреляция

отрицательная. Чем ближе r_{xy} к ± 1, тем уже эллипс и тем теснее экспериментальные значения группируются около прямой линии.



а) б)

Рис. 9. Положительная и отрицательная корреляция

4) Здесь же следует обратить внимание на то, что линия, вдоль которой группируются точки, может быть не только прямой, а иметь любую другую форму: парабола, гипербола и т. д. В этих случаях мы рассматривали бы так называемую, нелинейную (или криволинейную) корреляцию.

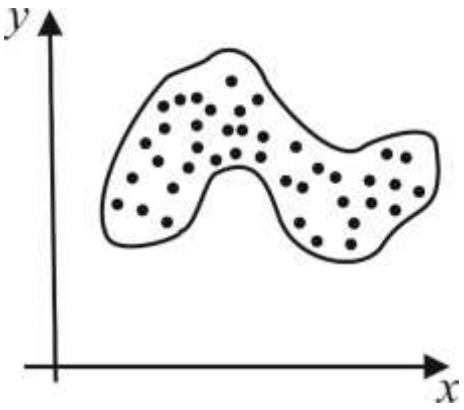


Рис. 9. Криволинейная корреляция

Пример.

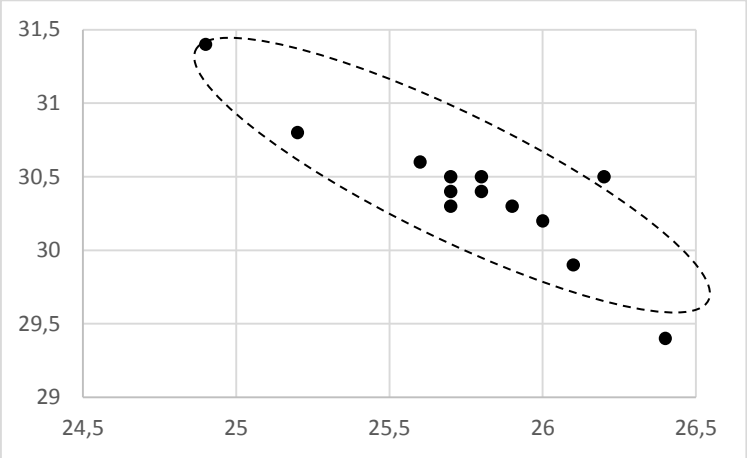
Определить форму и направление взаимосвязи между результатами в беге на первой и второй половине дистанции 400 м у 13 исследуемых с помощью построения графика корреляционного поля, если данные выборок таковы:

x_i , с ~ 25,2; 26,4; 26,0; 25,8; 24,9; 25,7; 25,7; 25,7; 26,1; 25,8; 25,9; 26,2; 25,6 (первые 200 м).

y_i , с ~ 30,8; 29,4; 30,2; 30,5; 31,4; 30,3; 30,4; 30,5; 29,9; 30,4; 30,3; 30,5; 30,6 (последние 200 м).

Решение.

Построим поле корреляции.



Точки попадают в область, ограниченную некоторым эллипсом, наклон влево. Можно предположить, что между величинами отрицательная корреляция.

3.2. Коэффициенты корреляции и их свойства

Коэффициент линейной корреляции указывает на ***степень выраженности тесноты связи*** между двумя переменными.

Коэффициент корреляции для генеральной совокупности, как правило, неизвестен, поэтому он оценивается по экспериментальным данным, представляющим собой выборку объема n пар значений (x_i, y_i), полученную при совместном измерении двух признаков X и Y.

Коэффициенты корреляции – удобный показатель связи, получивший широкое применение в практике. К их основным свойствам необходимо отнести следующие:

1. Коэффициенты корреляции способны характеризовать только линейные связи, т. е. такие, которые выражаются уравнением линейной функции. О криволинейной связи с их помощью ничего сказать нельзя. При наличии нелинейной зависимости между варьирующими признаками следует использовать другие показатели связи.

2. Значения коэффициентов корреляции есть безразмерная величина, которая не может быть меньше -1 и больше +1, т. е. значения коэффициентов корреляции – это отвлеченные числа, лежащее в пределах $-1 \le r_{xy} \le 1$.

3. При независимом варьировании признаков, когда связь между ними отсутствует, $r_{xy} = 0$.

4. При положительной, или прямой, связи, когда с увеличением значений одного признака возрастают значения другого, коэффициент корреляции приобретает положительный (+) знак и находится в пределах от 0 до +1, т. е. $0 < r_{xy} < 1$.

5. При отрицательной, или обратной, связи, когда с увеличением значений одного признака соответственно уменьшаются значения другого, коэффициент корреляции сопровождается отрицательным (–) знаком и находится в пределах от 0 до –1, т. е. $-1 < r_{xy} < 0$.

6. Чем сильнее связь между признаками, тем ближе величина коэффициента корреляции к |1|. Если $r_{xy} = \pm 1$, то корреляционная связь переходит в функциональную, т. е. каждому значению признака *X* будет соответствовать одно или несколько строго определенных значений признака *Y*.

7. Только по величине коэффициентов корреляции нельзя судить о достоверности корреляционной связи между признаками. Этот параметр зависит от числа степеней свободы $k = n - 2$, где: n – число коррелируемых пар показателей *X* и *Y*. Чем больше n, тем выше достоверность связи при одном и том же значении коэффициента корреляции.

В практической деятельности, когда число коррелируемых пар признаков *X* и *Y* не велико ($n \leq 30$), то при оценке зависимости между показателями используется используются некоторые ***условные*** границы коэффициента корреляции, которые служат для интерпретации полученных на выборке значений.

Общая классификация корреляционных связей

(по Ивантер Э.В., Коросову А.В., 1992):

- **сильная**, или **тесная** при коэффициенте корреляции $|r| \geq 0{,}70$;
- **средняя** при $0{,}50 \leq |r| < 0{,}7$;
- **умеренная** при $0{,}30 \leq |r| < 0{,}5$;
- **слабая** при $0{,}20 \leq |r| < 0{,}3$;
- **очень слабая** при $|r| \leq 0{,}2$.

В общем случае, если $|r_{xy}|\sqrt{n-1} > 3$, т о связь случайных величин *X* и *Y* достаточно вероятна.

Таким образом, коэффициент корреляции может служить мерой ли-

нейной взаимосвязи двух случайных (т. е. изменчивых) величин. Для вычисления коэффициента корреляции ***надо знать совместное распределение*** двух величин (показателей). Часто бывает, что на практике такого распределения мы достоверно не знаем (у нас может не быть достаточной информации, необходимых сведений). Однако можно провести измерения в ходе ***выборочного*** исследования и тогда для каждого объекта исследования зафиксировать значения X_i и Y_i и по выборке получить сведения о совместном распределении X и Y**.** В результате будут получены пары наблюдений

x_1	x_2	x_3	…	x_n
y_1	y_2	y_3	…	y_n

по которым можно ***оценить*** (т. е. найти ***примерное*** значение) теоретический (истинный, характерный для изучаемой генеральной совокупности) коэффициент корреляции.

Пусть дана система связь СВ X и Y. Пусть в результате n испытаний получено n точек $(x_1;y_1),(x_2;y_2),...,(x_n;y_n)$. Пусть требуется вычислить коэффициент корреляции этой системы СВ. Приняв во внимание закон больших чисел можно заменить $M(X)$ и $M(\mathrm{Y})$ средними арифметическими соответствующих СВ. Имеют место следующие приближенные равенства:

$$M(X)\approx\bar{x}=\frac{1}{n}\sum_{i=1}^{n}x_i n_i,\qquad M(Y)\approx\bar{y}=\frac{1}{n}\sum_{i=1}^{n}y_i n_i$$

$$\sigma_x^{\,2}\approx\frac{1}{n}\sum_{i=1}^{n}x_i^{\,2}n_i-(\bar{x})^2,\qquad \sigma_y^{\,2}\approx\frac{1}{n}\sum_{i=1}^{n}y_i^{\,2}n_i-(\bar{y})^2$$

$$COV(X,Y)\approx\frac{1}{n}\sum_{i=1}^{n}x_i y_i n_i-\bar{x}\cdot\bar{y}$$

$$r_{xy}=\frac{COV(X,Y)}{\sigma_x\sigma_y}$$

Пример:

В результате 79 опытов получена таблица

Y / X	0,5	0,6	0,7	0,8
0,5	0	2	0	8
0,6	0	4	2	9
0,7	2	12	3	1
0,8	21	14	0	0
0,9	1	0	0	0

Определить коэффициент корреляции.

Решение:

$$\bar{x} = \frac{0{,}5 \cdot 10 + 0{,}6 \cdot 15 + 0{,}7 \cdot 18 + 0{,}8 \cdot 35 + 0{,}9 \cdot 1}{79} = \frac{55{,}5}{79} = 0{,}703$$

$$\bar{y} = \frac{0{,}5 \cdot 24 + 0{,}6 \cdot 32 + 0{,}7 \cdot 5 + 0{,}8 \cdot 18}{79} = 0{,}622$$

$$\sigma_x^2 = \frac{0{,}5^2 \cdot 10 + 0{,}6^2 \cdot 15 + 0{,}7^2 \cdot 18 + 0{,}8^2 \cdot 35 + 0{,}9^2 \cdot 1}{79} - 0{,}703^2 =$$

$$= 0{,}505 - 0{,}703^2 = 0{,}012$$

$$\sigma_x = \sqrt{0{,}012} = 0{,}11$$

$$\sigma_y^2 = \frac{0{,}5^2 \cdot 24 + 0{,}6^2 \cdot 32 + 0{,}7^2 \cdot 5 + 0{,}8^2 \cdot 18}{79} - 0{,}622^2 = 0{,}012$$

$$\sigma_y = \sqrt{0{,}012} = 0{,}108$$

$$COV(X, Y) = \frac{1}{79}(0{,}5 \cdot 2 \cdot 0{,}6 + 0{,}5 \cdot 8 \cdot 0{,}8 + 0{,}6 \cdot 4 \cdot 0{,}6 + 0{,}6 \cdot 2 \cdot 0{,}7 + 0{,}6 \cdot 9 \cdot 0{,}8 +$$

$$+ 0{,}7 \cdot 2 \cdot 0{,}5 + 0{,}7 \cdot 12 \cdot 0{,}6 + 0{,}7 \cdot 3 \cdot 0{,}7 + 0{,}7 \cdot 1 \cdot 0{,}8 + 0{,}8 \cdot 21 \cdot 0{,}5 + 0{,}8 \cdot 14 \cdot 0{,}6 +$$

$$+ 0{,}9 \cdot 1 \cdot 0{,}5) - 0{,}703 \cdot 0{,}622 = -0{,}01$$

$$r_{xy} = \frac{-0{,}01}{0{,}11 \cdot 0{,}108} = -0{,}842$$

Между величинами обратная тесная связь.

3.3. Линейная корреляция

Наиболее полно в статистике разработана методология парной корреляции, рассматривающей влияние вариации одного факторного признака на вариацию результативного. Исследование парной корреляции осуществляется на основе корреляционного анализа, который предполагает последовательное решение ряда задач:

- выявление связи;
- описание связи в табличной и графической формах;
- измерение тесноты связи;
- формулировка выводов о характере существующей связи.

Описание выявленной связи при проведении корреляционного анализа проводится в двух формах – табличной и графической. При табличном описании связи статистические единицы группируются по значению факторного признака (располагаются в порядке его возрастания или убывания).

3.3.1. Эмпирическая линия регрессии

Графическое описание связи заключается в построении линии эмпирической регрессии – ломаной линии, соединяющей на корреляционном поле точки, абсциссами которых являются индивидуальные (групповые) значения факторного признака, а ординатами – соответствующие (средние) значения результативного признака. Линия эмпирической регрессии отражает основную тенденцию рассматриваемой зависимости. Если по своему виду она приближается к прямой линии, то можно предположить наличие прямолинейной связи между признаками.

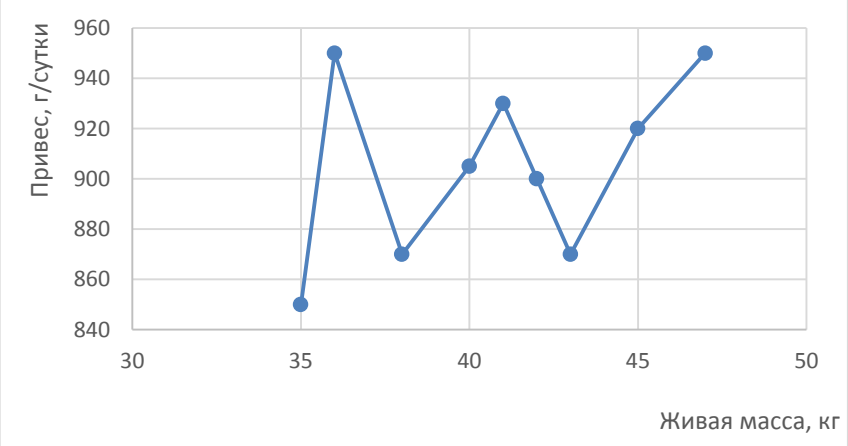
Пример.

Построить эмперическую линию регрессии Y к X по данным, представленным в таблице.

Живая масса, кг – *x*	**Привес, г/сутки – *y***
35	850
36	950
38	870
40	905
41	930
42	900
43	870
45	920
47	950

Решение.

На горизонтальной оси x системы координат отметим значения независимой переменной. На вертикальной оси y – значения зависимой переменной, соответствующие значениям независимой переменной. Соединяющая все точки линия представляет собой эмпирическую линию регрессии Y по X.



Пример.

Построить эмпирическую линию регрессии Y к X по данным, представленным в таблице.

Прибыль, млн. руб., y	Выпуск продукции, млн. руб., x				
	47	59	71	83	95
13,35	3	2			
15,85	1	3	11	2	
18,35			1	4	3

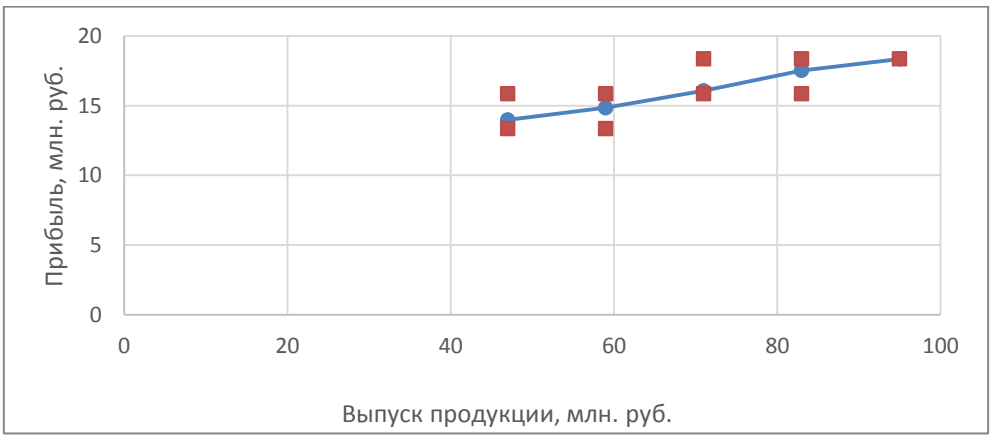
Решение.

В данном случае эмпирическая линия регрессии Y к X – это ломаная, соединяющая точки $(x_i; \bar{y}_i)$.

Вычислим средние значения результативного признака для каждого значения независимой переменной.

x	n_x	$\bar{y}_x$
47	4	$\bar{y}_{x=47} = \frac{13{,}35 \cdot 3 + 15{,}85 \cdot 1}{4} = 13{,}98$
59	5	$\bar{y}_{x=59} = \frac{13{,}35 \cdot 2 + 15{,}85 \cdot 3}{5} = 14{,}85$
71	12	$\bar{y}_{x=71} = \frac{15{,}85 \cdot 11 + 18{,}35 \cdot 1}{12} = 16{,}06$
83	6	$\bar{y}_{x=71} = \frac{15{,}85 \cdot 2 + 18{,}35 \cdot 4}{6} = 17{,}52$
95	3	$\bar{y}_{x=71} = \frac{18{,}35 \cdot 3}{3} = 18{,}35$

Построенная *ломаная* проходит максимально близко к точкам корреляционного поля, при этом учитываются весомость частот n_{ij}, на основе которых были вычислены значения $\bar{y}_i$.



3.3.2. Выборочное уравнение прямой линии регрессии

В практических исследованиях возникает необходимость **аппроксимировать** (математически описать приблизительно) корреляционную зависимость между двумя признаками уравнением. Для линейной зависимости сделать это относительно просто: вытянутое корреляционное поле заменить усредненной прямой линией и найти ее уравнение по статистическим данным коррелируемых признаков. В прямоугольной системе координат уравнение прямой линии записывается в виде: $\hat{y}_x = a_1 x + b_1$ или $\hat{x}_y = a_2 y + b_2$ (см. рис. 10).

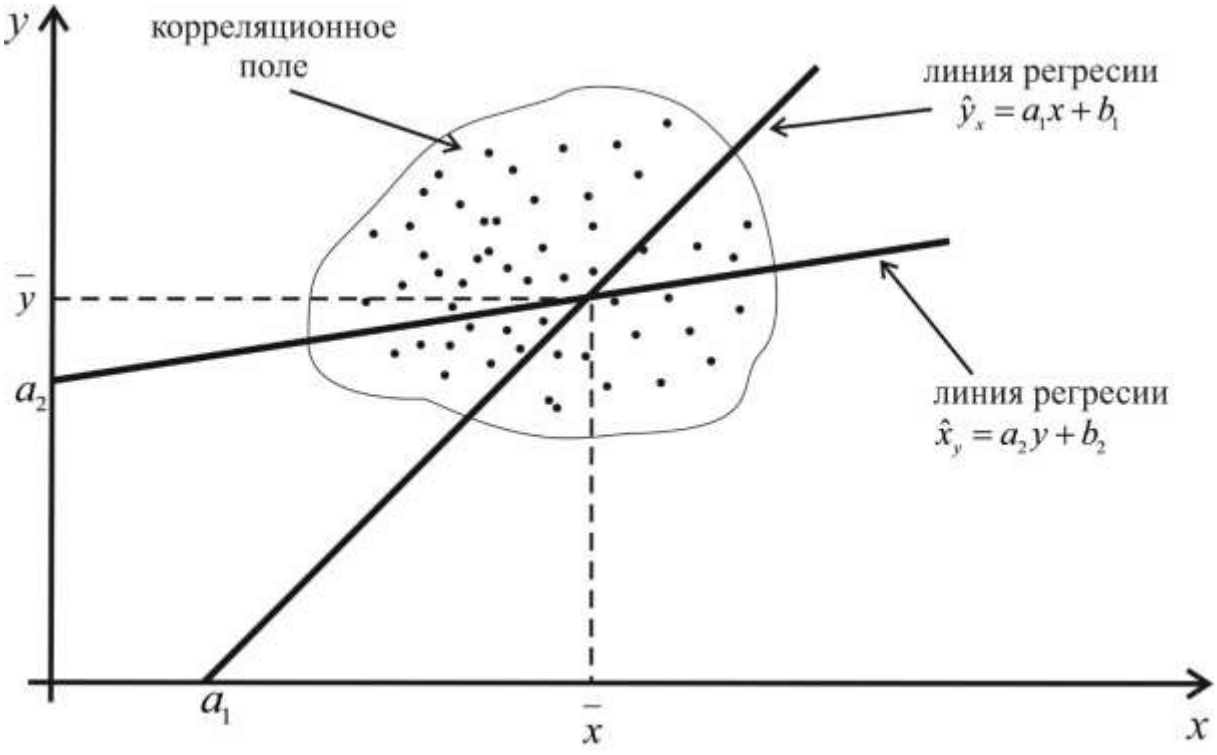


Рис. 10. Регрессионный анализ

Это математическое выражение корреляционной зависимости называется **уравнением регрессии.** Коэффициенты *a* и *b* называются **параметрами уравнения регрессии.** Параметр *b* определяет на графике отрезок, отсекаемый графиком уравнения (прямой линией) на оси *Y* (*X*). Параметр *a* показывает, как изменяется признак *Y*(*X*) при изменении признака *X (Y)*.

Это "*a*" еще называют коэффициентом регрессии обозначают $\rho_{yx} = r_{xy}\frac{\sigma_y}{\sigma_x}$ или $\rho_{xy} = r_{xy}\frac{\sigma_x}{\sigma_y}$.

Выборочное уравнение прямой линии регрессии *Y* на *X* имеет вид:

$$\hat{y}_x - \bar{y} = r_{xy}\frac{\sigma_y}{\sigma_x}(x - \bar{x})$$

$\bar{x}, \bar{y}$ – выборочные средние признаков *X* и *Y*, σ_x, σ_y – выборочные средние квадратические отклонения, r_{xy} – выборочный коэффициент корреляции.

Выборочное уравнение прямой линии регрессии *X* на *Y* имеет вид:

$$\hat{x}_y - \bar{x} = r_{xy}\frac{\sigma_x}{\sigma_y}(y - \bar{y})$$

Замечания:

1. Если выборка имеет достаточно большой объем и хорошо представляет генеральную совокупность (репрезентативна), то заключение о тесноте линейной зависимости между признаками, полученное по данным выборки, в известной степени может быть распространено и на генеральную совокупность. Например, для оценки коэффициента корреляции r_{xy}, нормально распределенной генеральной совокупности (при $n > 50$) можно воспользоваться формулой

$$r_{xy} - 3\cdot\frac{1-r_{xy}^2}{\sqrt{n}} \le r_{\Gamma} \le r_{xy} - 3\cdot\frac{1+r_{xy}^2}{\sqrt{n}}$$

2. Знак выборочного коэффициента корреляции совпадает со знаком выборочных коэффициентов регрессии что следует из формул $\rho_{yx} = r_{xy}\frac{\sigma_y}{\sigma_x}$ и $\rho_{xy} = r_{xy}\frac{\sigma_x}{\sigma_y}$.

3. Выборочный коэффициент корреляции равен среднему геометрическому выборочных коэффициентов регрессии: $\rho_{yx}\cdot\rho_{xy} = r_{xy}^2$

Пример.

Для примера п. 3.2. определить уравнения линий регрессии.

$|r_{xy}|\sqrt{n-1} = 0{,}867\sqrt{79-1} = 7{,}66 > 3$, связь достаточно вероятна.

Выборочное уравнение прямой линии регрессии *Y* на *X*

$$\hat{y}_x - 0{,}622 = -0{,}842\frac{0{,}108}{0{,}11}(x - 0{,}703)$$

$$\overline{\hat{y}_x} = -0{,}827x + 1{,}203$$

Выборочное уравнение прямой линии регрессии X на Y

$$\hat{x}_y - 0{,}703 = -0{,}842\frac{0{,}11}{0{,}108}(y - 0{,}622)$$

$$\hat{x}_y = -0{,}857y + 1{,}236$$

Если данные наблюдений над признаками X и Y заданы в виде таблицы с равноотстоящими вариантами, то переходят к условным вариантам:

$$u_i = \frac{x_i - C_1}{h_1}, v_j = \frac{y_j - C_2}{h_2}$$

где C_1 – "ложный нуль" вариант X, C_2 – "ложный нуль" вариант Y. В качестве ложного нуля выгодно принять варианту, которая расположена примерно в середине вариационного ряда или варианту, имеющую наибольшую частоту. h_1 – шаг X, h_2 – шаг Y.

В этом случае

$$r_{uv} = \frac{\sum n_{uv}uv - n\bar{u}\cdot\bar{v}}{n\sigma_u\sigma_v}$$

$$\bar{u} = \frac{\sum n_u u}{n}, \bar{v} = \frac{\sum n_v v}{n}, \sigma_u = \sqrt{\overline{u^2} - (\bar{u})^2}, \sigma_v = \sqrt{\overline{v^2} - (\bar{v})^2}$$

$$\bar{x} = \bar{u}h_1 + C_1,\ \bar{y} = vh_2 + C_2,\ \sigma_x = \sigma_u h_1,\ \sigma_y = \sigma_v h_2$$

Для расчетов удобно использовать расчетные таблицы. Рассмотрим на примере.

Пример.

Найти выборочные уравнения прямой линии регрессии Y на X по данным в таблице:

Y	X				
	20	25	30	35	40
16	4	6	–	–	–
26	-	8	10	–	–
36	-	-	32	3	9
46	-	-	4	12	6
56	-	-	-	1	5

Перейдем к условным вариантам. В качестве ложных нулей возьмем: $C_1 = 30$, $C_2 = 36$. По условию $h_1 = 5$, $h_2 = 10$. Тогда:

$$u_i = \frac{x_i - 30}{5}, v_j = \frac{y_j - 36}{10}$$

Составим расчетную таблицу.

u \ v	-2	-1	0	1	2	n_v	$n_v v$	$n_v v^2$	$n_{uv}uv$
-2	4	6	-	-	-	10	-20	40	28
-1	-	8	10	-	-	18	-18	18	8
0	-	-	32	3	9	44	0	0	0
1	-	-	4	12	6	22	22	22	24
2	-	-	-	1	5	6	12	24	22
n_u	4	14	46	16	20	Σ100	-4	104	82
$n_u u$	-8	-14	0	16	40	34	-	-	-
$n_u u^2$	16	14	0	16	80	126	-	-	-
$n_{uv}uv$	16	20	0	14	32	82	-	-	-

Совпадение сумм последнего столбца и последней строки свидетельствует о правильности вычислений.

Используя суммы по столбцам и по строкам находим числовые характеристики для условных вариант:

$$\bar{u} = \frac{34}{100} = 0{,}34;\ \ \bar{v} = \frac{-4}{100} = -0{,}04$$

$$\overline{u^2} = \frac{126}{100} = 1{,}26;\ \ \overline{v^2} = \frac{104}{100} = 1{,}04$$

$$\sigma_u = \sqrt{\overline{u^2} - \left(\bar{u}\right)^2} = \sqrt{1{,}26 - \left(0{,}34\right)^2} = 1{,}07$$

$$\sigma_v = \sqrt{\overline{v^2} - \left(\bar{v}\right)^2} = \sqrt{1{,}04 - \left(-0{,}04\right)^2} = 1{,}02$$

$$r_{uv} = \frac{\sum n_{uv}uv - n\bar{u}\cdot\bar{v}}{n\sigma_u\sigma_v} = \frac{82 - 100\cdot 0{,}34\cdot\left(-0{,}04\right)}{100\cdot 1{,}07\cdot 1{,}02} = 0{,}76$$

Переходим к первоначальным вариантам:

$$\bar{x} = \bar{u}h_1 + C_1 = 0{,}34\cdot 5 + 30 = 31{,}70,$$

$$\bar{y} = vh_2 + C_2 = -0{,}04\cdot 10 + 36 = 35{,}60$$

$$\sigma_x = \sigma_u h_1 = 5\cdot 1{,}07 = 5{,}35$$

$\sigma_y = \sigma_v h_2 = 10 \cdot 1{,}02 = 10{,}2$

$r_{xy} = 0{,}76$

$\hat{y}_x - 35{,}60 = 0{,}76 \cdot \frac{10{,}2}{5{,}35}(x - 31{,}70)$

$\hat{y}_x = 1{,}45x - 10{,}36$ – искомое уравнение регрессии.

3.4. Криволинейная корреляция

Если график регрессии – кривая линия, то корреляцию называют криволинейной. В частности, в случае параболической корреляции второго порядка выборочное уравнение регрессии *Y* на *X* имеет вид

$$\hat{y}_x = Ax^2 + Bx + C$$

Неизвестные параметры *A*, *B* и *C* находят (например, методом Гаусса) из системы уравнений:

$$\begin{cases} \left(\sum n_x x^4\right)A + \left(\sum n_x x^3\right)B + \left(\sum n_x x^2\right)C = \sum n_x \bar{y}_x x^2 \\ \left(\sum n_x x^3\right)A + \left(\sum n_x x^2\right)B + \left(\sum n_x x\right)C = \sum n_x \bar{y}_x x \\ \left(\sum n_x x^2\right)A + \left(\sum n_x x\right)B + nC = \sum n_x \bar{y}_x \end{cases}$$

Аналогично находится выборочное уравнение регрессии *X* на *Y*.

$$\hat{x}_y = Ay^2 + By + C$$

Для оценки силы корреляции *Y* на *X* служит *выборочное корреляционное отношение* (отношение межгруппового среднего квадратического к общему среднему отклонения признака *Y*:

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

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

Диапазон изменения этого показателя: $0 \le \eta \le 1$. Нулевое значение эмпирического корреляционного отношения означает отсутствие связи между результативным и факторным признаками, при η = 1 связь классифицируется как функциональная. Используя численное значение эмпирического корреляционного соотношения, связь можно классифицировать по шкале Чеддока.

Шкала Чеддока

η	**Характеристика связи**
$0 \le \eta < 0{,}1$	отсутствует
$0{,}1 \le \eta < 0{,}3$	слабая
$0{,}3 \le \eta < 0{,}5$	умеренная
$0{,}5 \le \eta < 0{,}7$	заметная
$0{,}7 \le \eta < 0{,}9$	тесная
$0{,}9 \le \eta < 0{,}99$	сильная
$0{,}99 \le \eta \le 1$	функциональная

Пример.

Найти выборочное уравнение регрессии по данным, приведенным в таблице. Оценить силу корреляционной связи.

Y	***X***		
	2	3	5
25	20		
45		30	1
110		1	48

Решение.

Составим расчетную таблицу.

x	n_x	$\bar{y}_x$	$n_x x$	$n_x x^2$	$n_x x^3$	$n_x x^4$	$n_x \bar{y}_x$	$n_x \bar{y}_x x$	$n_x \bar{y}_x x^2$
2	20	25	40	80	160	320	500	1000	2000
3	31	47,1	93	279	837	2511	1460	4380	13141
5	49	108,67	248	1225	6125	30625	5325	26624	133121
Σ	100		378	1584	7122	33456	7285	32004	148262

Подставив числа, содержащиеся в последней строке таблице, получим систему уравнений относительно неизвестных коэффициентов *A, B, C:*

$$\begin{cases} 33456A + 7122B + 1584C = 148262 \\ 7122A + 1584B + 378C = 32004 \\ 1584A + 378B + 100C = 7285 \end{cases}$$

Решив эту систему, найдем: A = 2,94; B = 7,27; C = - 1,25. Получим уравнение регрессии: $\hat{y}_x = 2{,}94x^2 + 7{,}27x - 1{,}25$.

Для вычисления выборочного корреляционного отношения найдем общую среднюю $\bar{y}$, общее среднее квадратическое отклонение σ_y и межгрупповое среднее квадратическое отклонение $\sigma_{межгр}$:

$$\bar{y} = \frac{\sum n_y y}{n} = \frac{20 \cdot 25 + 31 \cdot 45 + 49 \cdot 110}{100} = 72{,}85$$

$$\sigma_y = \sqrt{\frac{\sum n_y (y - \bar{y})^2}{n}} = \sqrt{\frac{20(25 - 72{,}85)^2 + 31(45 - 72{,}85)^2 + 49(110 - 72{,}85)^2}{100}} = 37{,}07$$

$$\sigma_{межгр} = \sqrt{\frac{\sum n_x \left(\hat{y}_x - \overline{y^2}\right)}{n}} =$$

$$= \sqrt{\frac{20(25 - 72{,}85)^2 + 31(47{,}1 - 72{,}85)^2 + 49(108{,}67 - 72{,}85)^2}{100}} = 35{,}95$$

Найдем искомое выборочное корреляционное отношение:

$$\eta_{yx} = \frac{\sigma_{межгр}}{\sigma_{общ}} = \frac{35{,}95}{37{,}07} = 0{,}97$$

По шкале Чеддока определяем, что между величинами сильная связь.

Вопросы для самоконтроля

1. Какая зависимость между признаками называется статистической? Приведите пример.

2. В чем отличие терминов "корреляционная связь" и "корреляционная зависимость"?

3. Сформулируйте основные задачи корреляционного анализа.

4. Сформулируйте свойства коэффициента корреляции Пирсона.

5. Какой вывод делает исследователь, если выборочный коэффициент корреляции Пирсона равен: 1) $r = -0{,}75$; 2) $r = 0{,}92$; 3) $r = 0{,}15$?

6. В чем состоит различие между функциональной и статистической зависимостью между случайными величинами?

7. Что следует сказать о зависимости двух случайных величин, если коэффициент корреляции равен нулю, единице?

8. Что такое регрессионный анализ?

9. Что такое эмпирическая простая линейная регрессия?

10. Запишите уравнения прямых регрессий *X* на *Y* и *Y* на *X*.

11. Как рассчитывается коэффициент корреляции?

12. Что представляет собой диаграмма рассеяния?

Задачи для самостоятельного решения

1. В таблице приведен ряд, устанавливающий связь между уровнем *IQ* и уровнем средней успеваемости студентов по математике.

X – уровень *IQ*	75	85	90	100	105	110	110	115	115	120	125	130	140
Y – средняя успеваемость	3,1	3,1	3,5	3,7	3,8	4,0	4,2	4,3	4,6	4,7	4,8	4,9	5,0

Существует ли взаимосвязь между уровнем *IQ* (признак X) и средним уровнем успеваемости по математике (признак *Y*)?

2. В результате независимых испытаний получены пары значений случайных величин *X* и *Y*:

xi	10	20	25	28	30
yi	4	8	7	12	14

Найти выборочное уравнение линейной регрессии и выборочный коэффициент корреляции. Построить прямые регрессии *Y* на *X* и *X* на *Y*.

3. Изучая зависимость между показателями *X* и *Y*, проведено обследование 10 объектов и получены следующие данные

x	120	70	100	55	75	85	110	80	60	95
y	4,6	2,6	4,3	2,4	3,1	3,8	4,2	2,9	2,7	3,4

Полагая, что между *X* и *Y* имеет место линейная корреляционная связь, определите выборочное уравнение регресии и выборочный коэффициент линейной регрессии. Построить диаграмму рассеяния и линию регресии. Сделать вывод о направлении и тесноте связи между показателями

4. Изучается зависимость себестоимости одного изделия (Y, р.) от величины выпуска продукции (X, тыс. шт.) по группе предприятий за отчетный период. Получены следующие данные. Провести корреляционно-регрессионный анализ зависимости себестоимости одного изделия от выпуска продукции.

X	2	3	4	5	6
Y	1,9	1,7	1,8	1,6	1,4 .

5. Найти выборочные уравнения прямых линий регрессии *Y* на *X* и *X* на *Y* по данным, приведенным в таблице:

Y	*X*						
	18	23	28	33	38	43	48
125		1					
150	1	2	5				
175		3	2	12			
200			1	8	7		
225					3	3	
250						1	1

6. Найти выборочные уравнения прямых линий регрессии Y на X и X на Y по данным, приведенным в таблице:

Y	X							
	5	10	15	20	25	30	35	40
100	2	1						
120	3	4	3					
140			5	10	8			
160				1		6	1	1
180							4	1

7. В таблице приведены данные обследования (количество человек) 20 мужчин возрастом 20, 40 и 50 лет (СВ_X – возраст, СВ_Y – число волос на голове в тыс. штук). Найти коэффициент корреляции между этими величинами, дать прогноз количества волос в возрасте 60 лет и спрогнозировать возраст полного облысения.

X / Y	46	38	30
20	6	2	-
40	3	4	1
50	-	1	3

4. СТАТИСТИЧЕСКАЯ ПРОВЕРКА ГИПОТЕЗ

4.1. Задачи статистической проверки гипотез

Пусть (x_1, x_2,…, x_n) – случайная выборка объема n из некоторой генеральной совокупности (конечной или бесконечной). Каждое значение x_i в этой выборке само является случайной величиной, даже если генеральная совокупность состоит из конечного числа элементов. Необходимо также иметь в виду, что случайная выборка из какой-либо генеральной совокупности должна соответствовать некоторой схеме испытаний, при реализации которой выявляется искомая случайная величина X. При этом полученные в вышеупомянутой серии испытаний значения случайной величины X должны быть *независимыми* и *распределены по тому же закону*, что и сама генеральная совокупность X (хотя бы и приближенно). *Статистической гипотезой* называется любое предположение относительно вида или параметров генерального распределения.

Статистической гипотезой называют гипотезу о виде неизвестного распределения генеральной совокупности или о параметрах известных распределений.

Методы математической статистики позволяют проверить предположения о законе распределения некоторой случайной величины (генеральной совокупности), о значениях параметров этого закона (например, *M(X)*, *D(X)*), о наличии корреляционной зависимости между случайными величинами, определенными на множестве объектов одной и той же генеральной совокупности.

По своему прикладному содержанию статистические гипотезы можно подразделить на несколько основных типов:

– о числовых значениях параметров;

– о равенстве числовых характеристик генеральных совокупностей;

– об однородности выборок;

– о согласии эмпирического распределения и выбранной модели;

– о стохастической независимости элементов выборки.

Статистическая гипотеза называется *параметрической*, если она содержит утверждение о значении конечного числа параметров распределения, которое считается известным.

Примеры параметрических статистических гипотез:

– нормально распределенная случайная величина X имеет математическое ожидание a и дисперсию σ^2;

– две нормально распределенные случайные величины имеют одинаковую дисперсию.

Непараметрическая гипотеза – это утверждение о виде распределения.

Например: – выборка (x_1, x_2,…, x_n) соответствует нормально распределенной случайной величине X.

Не располагая сведениями о всей генеральной совокупности, высказанную гипотезу сопоставляют по определенным правилам с выборочными данными и делают вывод о том, можно принять гипотезу или нет. Эта процедура сопоставления называется *проверкой гипотезы.*

Проверить статистическую гипотезу – это значит проверить, согласуются ли данные, полученные из выборки с этой гипотезой.

Гипотезу, выдвинутую для проверки ее согласия с выборочными данными, называют **нулевой гипотезой** и обозначают H_0. Вместе с гипотезой

H_0 выдвигается **альтернативная** или **конкурирующая** гипотеза, которая обозначается H_1.

H_0 и H_1 – две взаимно исключающие гипотезы.

Отметим, что для одной основной гипотезы может быть выдвинуты несколько альтернативных.

Так, например, пусть случайная величина X имеет нормальное распределение с математическим ожиданием a и дисперсией σ^2. Рассмотрим основную гипотезу:

$H_0 : a = 0, \sigma^2 = 1$.

В качестве альтернативных могут быть выдвинуты, например, такие гипотезы:

1) $H_1 : a = 0, \sigma^2 = 2$;

2) $H_1 : a \neq 0, {}^2 = 1$.

Можно было бы выдвинуть альтернативные гипотезы.

H_1: $a < 0$ (так называемая *левосторонняя* гипотеза) или H_1: $a > 0$ (*правосторонняя* гипотеза).

Простой называют гипотезу, содержащую только одно предположение, **сложной** – гипотезу, состоящую из конечного или бесконечного числа простых гипотез.

Пример. Для показательного распределения гипотеза H_0: $\lambda = 2$ – простая, H_0: $\lambda > 2$ – сложная, состоящая из бесконечного числа простых (вида $\lambda = c$, где c – любое число, большее 2).

Для проверки статистической гипотезы используется специально подобранная случайная величина K с известным законом распределения, называемая **статистическим критерием**. Этот критерий называют еще **критерием согласия** (имеется в виду согласие принятой гипотезы с результатами, полученными из выборки).

Множество ее возможных значений разбивается на два непересекающихся подмножества: одно из них (*критическая область*) содержит значения критерия, при которых нулевая гипотеза отклоняется, второе (*область принятия гипотезы*) – значения K, при которых она принимается. Значения K, отделяющие критическую область от области принятия гипотезы, называются *критическими точками* $k_{кр}$.

Наблюдаемым значением $K_{набл}$ называют значение критерия, вычисленное по выборкам.

Критической областью называют область значений критерия, при которых нулевую гипотезу отвергают, **областью принятия гипотезы** – область значений критерия, при которых 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.

Критерия, позволяющего точно (на 100 %) узнать, верна гипотеза H_0 или нет, не существует в силу ограниченности и случайности выборки (выборка не содержит всей информации о генеральной совокупности). Абсолютно точную информацию о генеральной совокупности мы бы получили, если бы исследовали *всю* генеральную совокупность. Мы исследуем *лишь выборку из нее*, поэтому не застрахованы от ошибки.

Отклоняя или принимая гипотезу H_0, можно допустить ошибку двух видов:

1) **Ошибка первого рода** совершается при отклонении гипотезы H_0 (т. е. принимается альтернативная H_1), тогда как на самом деле гипотеза H_0 верна. Вероятность ошибки первого рода называется **уровнем значимости** $\alpha = P(H_1 / H_0)$**.**

2) **Ошибка второго рода** совершается при принятии гипотезы H_0, тогда как на самом деле высказывание H_0 неверно и следовало бы принять гипотезу H_1; вероятность ошибки второго рода обозначим как $\beta = P(H_0 / H_1)$. Величина $(1 - \beta)$ – **мощность критерия.**

Решение, принимаемое о H_0 по выборке	Гипотеза H_0 отвергается, принимается H_1	Гипотеза H_0 принимается
Гипотеза H_0 верна	Ошибка 1-го рода, ее вероятность α	Правильное решение, его вероятность $1 - \alpha$
Гипотеза H_0 неверна, верна H_1	Правильное решение, его вероятность $1 - \beta$	Ошибка 2-го рода, его вероятность β

Чем выше мощность, тем меньше вероятность совершить ошибку второго рода. Чем меньше будут ошибки первого и второго рода, тем точнее статистический вывод. Однако при заданном объеме выборки одновременно уменьшить α и β невозможно. Единственный способ одновременного уменьшения α и β состоит в увеличении объема выборки.

4.2. Отыскание критических областей

Пусть случайная величина K – статистический критерий проверки некоторой гипотезы H_0.

Выберем некоторую малую вероятность α – уровень значимости.. Определим **критическое значение критерия** $k_{кр}$ как решение одного из трех уравнений, в зависимости от вида нулевой и конкурирующей гипотез:

$$P\,(K > k_{кр}) = \alpha \qquad (1)$$

$$P\,(K < k_{кр}) = \alpha \qquad (2)$$

$$P\,((K < k_{кр1}) \cap (K > k_{кр2})) = \alpha \qquad (3)$$

Решение уравнения (1) заключается в следующем: по вероятности α, зная функцию $p_K(x)$, заданную как правило таблицей, нужно определить $k_{кр}$. Если гипотеза H_0 справедлива, то вероятность того, что критерий K превзойдет некоторое значение $k_{кр}$ очень мала – 0,05, 0,01 или еще меньше, в зависимости от нашего выбора. Если $K_{набл}$ – значение критерия K, рассчитанное по выборочным данным, превзошло значение $k_{кр}$, это означает, что выборочные данные не дают основания для принятия нулевой гипотезы H_0 (например, если α=0,01 , то можно сказать, что произошло событие, которое при справедливости гипотезы H_0 встречается в среднем не чаще, чем в одной из ста выборок). В этом случае говорят, что **гипотеза H_0 не согласуется с выборочными данными и должна быть отвергнута**. Если $K_{набл}$ не превосходит $K_{кр}$, то говорят, что **выборочные данные не противоречат гипотезе H_0**, и нет оснований отвергать эту гипотезу.

Уравнение (1) определяет **правостороннюю критическую область**. Область $K > K_{кр}$ – критическая область. Если значение $K_{набл}$ попадает в критическую область, то гипотеза H_0 отвергается. Область $K < K_{кр}$ – областью принятия гипотезы. Если значение $K_{набл}$ попадает в область принятия гипотезы, то гипотеза H_0 принимается. Рисунок 11. иллюстрирует решение уравнения (1). Здесь $p_K(x)$ – известная плотность распределения случайной величины K при условии справедливости гипотезы H_0.

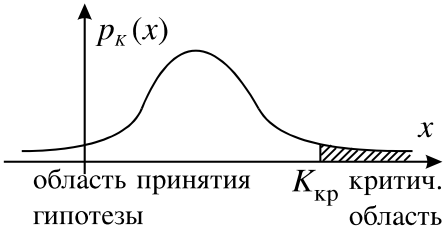


Рис. 11. Правосторонняя критическая область

Уравнение (2) определяет **левостороннюю критическую область**. Ее изображение приводится на рисунке 12.

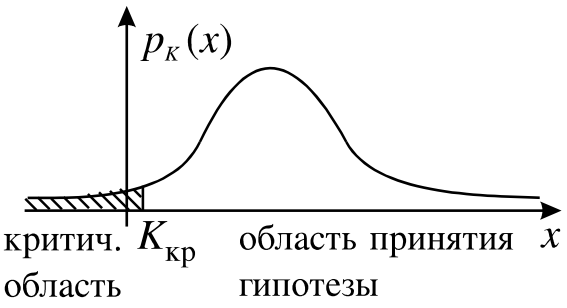


Рис. 12. Левосторонняя критическая область

Уравнение (3) определяет **двустороннюю критическую область**. Такая область изображена на рисунке 13. Здесь критическая область состоит из двух частей. В случае двусторонней критической области границы ее частей $k_{кр1}$ и $k_{кр2}$ определяются таким образом, чтобы выполнялось условие:

$$P(K \le K_{кр}) = P(K \ge K_{кр}) = \alpha / 2.$$

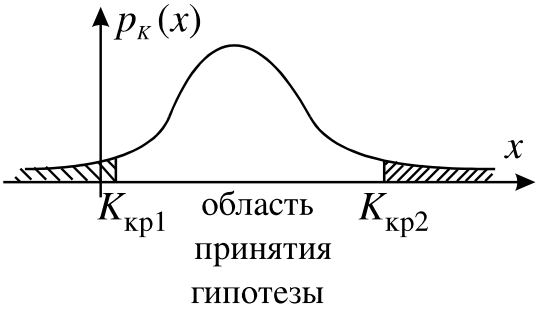


Рис. 13. Двусторонняя критическая область

Итак, процесс проверки гипотезы состоит из следующих этапов:

1) выбирается статистический критерий K;

2) вычисляется его наблюдаемое значение $K_{набл}$ по имеющейся выборке;

3) поскольку закон распределения K известен, определяется (по известному уровню значимости α) критическое значение $k_{кр}$, разделяющее критическую область и область принятия гипотезы (например, если $P(K > k_{кр}) = \alpha$, то справа от $k_{кр}$ располагается критическая область, а слева – область принятия гипотезы);

4) если вычисленное значение $K_{набл}$ попадает в область принятия гипотезы, то нулевая гипотеза принимается, если в критическую область – нулевая гипотеза отвергается.

4.3. Сравнение двух дисперсий нормальных генеральных совокупностей

На практике задача сравнения дисперсий возникает, если требуется сравнить точность приборов, инструментов, самих методов измерений и т. д. Очевидно, предпочтительнее тот прибор, инструмент и метод, которые обеспечивают наименьшее рассеяние измерений, т. е. наименьшую дисперсию.

Пусть генеральные совокупности X и Y распределены нормально. По независимым выборкам объемов n_1 и n_2, извлеченным из этих совокупностей, найдены исправленные выборочные дисперсии s_x^2 и s_y^2.. Требуется по исправленным дисперсиям при заданном уровне значимости α проверить нулевую гипотезу, состоящую в том, что генеральные дисперсии рассматриваемых совокупностей равны между собой: H_0: $D(X) = D(Y)$.

Учитывая, что исправленные дисперсии являются несмещенными оценками генеральных дисперсий, т. е. $M\left(s_x^2\right) = D(X)$, $M\left(s_y^2\right) = D(Y)$, нулевую гипотезу можно записать так:

$$H_0 : M\left(s_x^2\right) = M\left(s_y^2\right).$$

Таким образом, требуется проверить, что математические ожидания исправленных выборочных дисперсий равны между собой. Такая задача ставится потому, что обычно исправленные дисперсии оказываются различными. Возникает вопрос: значимо (существенно) или незначимо различаются исправленные дисперсии?

Если окажется, что нулевая гипотеза справедлива, т. е. генеральные дисперсии одинаковы, то различие исправленных дисперсий незначимо и объясняется случайными причинами, в частности, случайным отбором объектов выборки. Например, если различие исправленных выборочных дисперсий результатов измерений, выполненных двумя приборами, оказалось незначимым, то приборы имеют одинаковую точность.

Если нулевая гипотеза будет отвергнута, т. е. генеральные дисперсии неодинаковы, то различие исправленных дисперсий значимо и не может быть объяснено случайными причинами, а является следствием того, что сами генеральные дисперсии различны. Например, если различие исправленных выборочных дисперсий результатов измерений, произведенных двумя приборами, оказалось значимым, то точность приборов различна.

В качестве критерия проверки нулевой гипотезы о равенстве генеральных дисперсий, примем отношение большей исправленной дисперсии к меньшей, т. е. случайную величину:

$$F = \frac{s_Б^2}{s_M^2}.$$

Величина F при условии справедливости нулевой гипотезы имеет распределение Фишера-Снедекора со степенями свободы $k_1=n_1$-1 и $k_2= n_2$-1, где n_1 – объем выборки, по которой вычислена большая исправленная дисперсия.

Критическая область строится в зависимости от вида конкурирующей гипотезы.

Первый случай.

Нулевая гипотеза H_0: $D(X) = D(Y)$.

Конкурирующая гипотеза H_1: $D(X) > D(Y)$.

В этом случае строят одностороннюю, а именно правостороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия F в эту область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости: $P\left(F > F_{кр}\left(\alpha; k_1; k_2\right)\right) = \alpha$.

Критическую точку $F_{кр}(\alpha, k_1, k_2)$ находят по таблице критических точек распределения Фишера-Снедекора, и тогда правосторонняя критическая область определяется неравенством $F > F_{кр}$, а область принятия нулевой гипотезы – неравенством $F < F_{кр}$.

Обозначим отношение большей исправленной дисперсии к меньшей, вычисленное по данным наблюдений, через $F_{набл}$ и сформулируем правило проверки нулевой гипотезы.

Правило 1.

1) Вычислить отношение большей исправленной дисперсии к меньшей, т. е.

$$F_{набл} = \frac{s_Б^2}{s_M^2}$$

2) По таблице критических точек распределения Фишера-Снедекора, по заданному уровню значимости и числам степеней свободы k_1 и k_2 (k_1 – число степеней свободы большей исправленной дисперсии) найти критическую точку $F_{кр}\ (\alpha, k_1, k_2)$.

3) Если $F_{набл} < F_{кр}$ – нет оснований отвергнуть нулевую гипотезу. Если $F_{набл} > F_{кр}$ – нулевую гипотезу отвергают.

Пример.

По двум независимым выборкам объемов n_1=12, n_2=15, извлеченным из нормальных генеральных совокупностей X и Y найдены исправленные выборочные дисперсии $s_x^2 = 11{,}41$; $s_y^2 = 6{,}52$. При уровне значимости α=0,05 проверить нулевую гипотезу H_0: $D(X)=D(Y)$ о равенстве генеральных дисперсий при конкурирующей гипотезе H_1: $D(X) > D(Y)$.

Решение.

$$F_{набл} = \frac{s_Б^2}{s_M^2} = \frac{11{,}41}{6{,}52} = 1{,}75.$$

Число степеней свободы: k_1 = 12-1=11; k_2 = 15-1=14

По таблице критических значении (Приложение…) находим $F_{кр}(0{,}05;11;14) = 2{,}56$.

Так как $F_{набл} < F_{кр}$ – нет оснований отвергнуть нулевую гипотезу о равенстве генеральных дисперсий.

Второй случай.

Нулевая гипотеза H_0: $D(X) = D(Y)$.

Конкурирующая гипотеза H_1: $D(X) \neq D(Y)$.

В этом случае строят двустороннюю критическую область исходя из требования, чтобы вероятность попадания критерия в эту область, в предположении справедливости нулевой гипотезы, была равна этому уровню значимости.

Наибольшая мощность (вероятность попадания критерия в критическую область при справедливости конкурирующей гипотезы) достигается тогда, когда вероятность попадания критерия в каждый из двух интервалов критической области равна $\alpha/2$.

Таким образом, если обозначить через F_1 левую границу критической области и через F_2 –правую, то должны иметь место соотношения (см. рисунок 14): $P(F < F_1) = \alpha/2$; $P(F > F_2) = \alpha/2$.

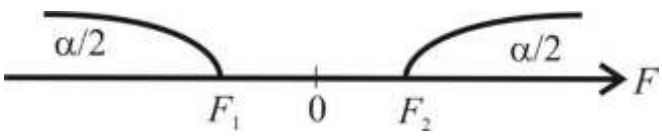


Рис. 14. Двусторонняя критическую область

Достаточно найти критические точки, чтобы найти саму критическую область: $F < F_1$, $F > F_2$, а также область принятия нулевой гипотезы: $F_1 < F < F_2$.

Правую критическую точку $F_2 = F_{кр}(\alpha/2; k_1; k_2)$ находят непосредственно по таблице критических точек распределения Фишера-Снедекора по уровню значимости и степеням свободы k_1 и k_2.

Левых критических точек эта таблица не содержит, и поэтому найти F_1 непосредственно по таблице невозможно. Достаточно найти правую критическую точку F_2 при уровне значимости, вдвое меньшем заданного. Тогда не только вероятность попадания критерия в "правую часть" критической области (т. е. правее F_2) равна $\alpha/2$, но и вероятность попадания этого критерия в "левую часть" критической области (т. е. левее F_1) будет также равна $\alpha/2$. Так как эти события несовместимы, то вероятность попадания рассматриваемого критерия во всю двустороннюю критическую область будет равна $\alpha/2$.

Таким образом, в случае конкурирующей гипотезы H_1: $D(X) \neq D(Y)$ достаточно найти критическую точку $F_2 = F_{кр}(\alpha/2, k_1, k_2)$.

Правило 2.

1) Вычислить отношение большей исправленной дисперсии к меньшей, т. е.

$$F_{набл} = \frac{s_Б^2}{s_М^2}$$

2) По таблице критических точек распределения Фишера-Снедекора по уровню значимости $\alpha/2$ (вдвое меньше заданного) и числам степеней свободы k_1 и k_2 (k_1 – число степеней свободы большей дисперсии) найти критическую точку $F_2 = F_{кр}(\alpha/2; k_1; k_2)$.

3) Если $F_{набл} < F_{кр}$ – нет оснований отвергать нулевую гипотезу.

Если $F_{набл} > F_{кр}$ – нулевую гипотезу отвергают.

Пример.

По двум независимым выборкам объемов n_1=12, n_2=15, извлеченным из нормальных генеральных совокупностей X и Y найдены исправленные выборочные дисперсии $s_x^2 = 1{,}23$; $s_y^2 = 0{,}41$. При уровне значимости α=0,01 проверить нулевую гипотезу H_0: $D(X) = D(Y)$ о равенстве генеральных дисперсий при конкурирующей гипотезе H_1: $D(X) \neq D(Y)$.

Решение.

$$F_{набл} = \frac{s_Б^2}{s_M^2} = \frac{1{,}23}{0{,}41} = 3.$$

По условию, конкурирующая гипотеза имеет вид H_1: $D(X) \neq D(Y)$, поэтому критическая область – двусторонняя.

Число степеней свободы: $k_1 = 10-1=9$; $k_2 = 18-1=17$

По таблице критических значении (Приложение 7) по уровню значимости вдвое меньше заданного находим $F_{кр}(0{,}05;9;17) = 2{,}5$.

Так как $F_{набл} > F_{кр}$, нулевую гипотезу о равенстве генеральных дисперсий отвергаем. Другими словами, выборочные исправленные дисперсии различаются значимо.

4.4. Сравнение двух средних нормальных генеральных совокупностей, дисперсии которых известны

Пусть генеральные совокупности X и Y распределены нормально, причем их дисперсии известны (например, из предшествующего опыта или найдены теоретически). По независимым выборкам объемов n и m ($n > 30$ и $m > 30$), извлеченным из этих совокупностей, найдены выборочные средние $\bar{x}$ и $\bar{y}$.

Требуется по выборочным средним при заданном уровне значимости α проверить нулевую гипотезу, состоящую в том, что генеральные средние (математические ожидания) рассматриваемых совокупностей равны между собой, т. е. H_0: $M(X) = M(Y)$.

Учитывая, что выборочные средние являются несмещенными оценками генеральных средних, т. е. $M(\bar{X}) = M(X)$ и $M(\bar{Y}) = M(Y)$, нулевую гипотезу можно записать так: $H_0; M(\bar{X}) = M(\bar{Y})$.

Таким образом, требуется проверить, что математические ожидания выборочных средних равны между собой. Такая задача ставится, потому что, как правило, выборочные средние являются различными. Возникает вопрос: значимо или незначимо различаются выборочные средние?

Если окажется, что нулевая гипотеза справедлива, т. е. генеральные средние одинаковы, то различие выборочных средних незначимо и объясняется случайными причинами и, в частности, случайным отбором объектов выборки.

Если нулевая гипотеза будет отвергнута, т. е. генеральные средние неодинаковы, то различие выборочных средних значимо и не может быть объяснено случайными причинами. А объясняется тем, что сами генеральные средние (математические ожидания) различны.

В качестве проверки нулевой гипотезы примем случайную величину

$$Z = \frac{\overline{X} - \overline{Y}}{\sigma\left(\overline{X} - \overline{Y}\right)} = \frac{\overline{X} - \overline{Y}}{\sqrt{D(X)/n + D(Y)/m}}.$$

Критерий Z – нормированная нормальная случайная величина, так как является линейной комбинацией нормально распределенных величин X и Y; сами эти величины распределены нормально как выборочные средние, найденные по выборкам, извлеченным из генеральных совокупностей; Z – нормированная величина, потому что $M(Z) = 0$, при справедливости нулевой гипотезы $D(Z) = 1$, поскольку выборки независимы.

Критическая область строится в зависимости от вида конкурирующей гипотезы.

Первый случай.

Нулевая гипотеза H_0: $M(X) = M(Y)$.

Конкурирующая гипотеза H_1: $M(X) \neq M(Y)$.

В этом случае строят двустороннюю критическую область исходя из требования, чтобы вероятность попадания критерия в эту область, в предположении справедливости нулевой гипотезы, была равна принятому уровню значимости α.

Наибольшая мощность критерия (вероятность попадания критерия в критическую область при справедливости конкурирующей гипотезы) достигается тогда, когда "левая" и "правая" критические точки выбраны так, что вероятность попадания критерия в каждый интервал критической области равна $\alpha/2$: $P\left(Z < Z_{лев.кр}\right) = \alpha/2$; $P\left(Z > Z_{прав.кр}\right) = \alpha/2$

Поскольку Z – нормированная нормальная величина, а распределение такой величины симметрично относительно нуля, критические точки симметричны относительно нуля.

Таким образом, если обозначить правую границу двусторонней критической области через $z_{кр}$, то левая граница $-z_{кр}$ (см. рис. 15).

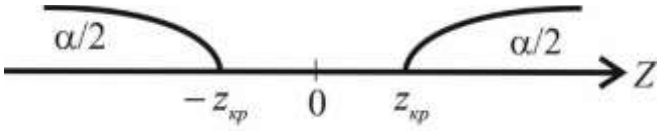


Рис. 15. Двусторонняя критическая область

Достаточно найти правую границу, чтобы найти саму двустороннюю критическую область $Z < -z_{кр}$, $Z > z_{кр}$ и область принятия нулевой гипотезы $(-z_{кр}, z_{кр})$.

Покажем как найти $z_{кр}$ – правую границу двусторонней критической области, пользуясь функцией Лапласа $\Phi(Z)$. Известно, что функция Лапласа определяет вероятность попадания нормированной нормальной случайной величины, например Z, в интервал $(0, z)$:

$$P(0 < Z < z) = \Phi(Z)$$

Так как распределение Z симметрично относительно нуля, то вероятность попадания Z в интервал $(0; +\infty)$ равна 0,5. Следовательно, если разбить этот интервал точкой $z_{кр}$, на интервалы $(0; z_{кр})$ и $(z_{кр}; 0)$, то по теореме сложения

$$P(0 < Z < z_{кр}) + P(Z > z_{кр}) = 0{,}5$$

Получаем $\Phi(z_{кр}) + \dfrac{\alpha}{2} = 0{,}5$

Следовательно, $\Phi(z_{кр}) = \dfrac{1-\alpha}{2}$.

Для того чтобы найти правую границу двусторонней критической области $z_{кр}$ достаточно найти значение аргумента функции Лапласа, которому соответствует значение функции, равное $\dfrac{1-\alpha}{2}$.

Тогда двусторонняя критическая область определяется неравенством $|Z| > z_{кр}$, а область принятия нулевой гипотезы неравенством $|Z| < z_{кр}$.

Правило 1.

1. Вычислить наблюдаемое значение критерия

$$Z_{набл} = \frac{\bar{x} - \bar{y}}{\sqrt{D(X)/n + D(Y)/n}}$$

2. По таблице функции Лапласа найти критическую точку по равенству $\Phi(z_{кр}) = \dfrac{1-\alpha}{2}$.

3. Если $|Z_{набл}| < z_{кр}$ – нет оснований отвергнуть нулевую гипотезу. Если $|Z_{набл}| > z_{кр}$ – нулевую гипотезу отвергают.

Пример.

По двум независимым выборкам объемов $n = 60$ и $m = 50$, извлеченным из нормальных генеральных совокупностей, найдены выборочные средние $\bar{x} = 1250$ и $\bar{y} = 1275$. Генеральные дисперсии известны: $D(X) = 120$, $D(Y)=100$. При уровне значимости 0,01, проверить нулевую гипотезу H_0: $M(X) = M(Y)$ при конкурирующей гипотезе H_1: $M(X) \neq M(Y)$.

Решение.

1) Найдем наблюдаемое значение критерия:

$$Z_{набл} = \frac{\bar{x} - \bar{y}}{\sqrt{D(X)/n + D(Y)/n}} = \frac{1250 - 1275}{\sqrt{\frac{120}{60} + \frac{100}{50}}} = -12{,}5$$

2) По условию конкурирующая гипотеза имеет вид H_1: $M(X) \neq M(Y)$, поэтому критическая область – двусторонняя. Найдем правую критическую точку по равенству

$$\Phi(z_{кр}) = \frac{1-\alpha}{2} = \frac{1-0{,}01}{2} = 0{,}495.$$

По таблице функции Лапласа (Приложение 2) находим $z_{кр} = 2{,}58$.

Так как $|Z_{набл}| > z_{кр}$ нулевую гипотезу отвергаем, выборочные средние различаются значимо.

Второй случай.

Нулевая гипотеза H_0: $M(X) = M(Y)$.

Конкурирующая гипотеза H_1: $M(X) > M(Y)$.

На практике такой случай имеет место, если профессиональные соображения позволяют предположить, что генеральная средняя одной совокупности больше генеральной средней другой. Например, если введено усовершенствование технологического процесса, то естественно допустить, что оно приведет к увеличению выпуска продукции.

В этом случае строят правостороннюю критическую область исходя из требования, чтобы вероятность попадания критерия в эту область, в предположении справедливости нулевой гипотезы, была равна принятому уровню значимости (см. рис. 16): $P(Z > Z_{кр}) = \alpha$.

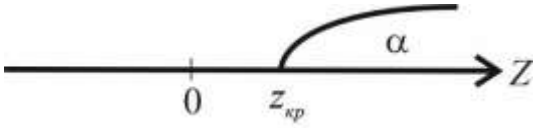


Рис. 16. Правосторонняя критическая область

Так как $P(0 < Z < z_{кр}) + P(Z > z_{кр}) = 0{,}5$ аналогично первому случаю получаем $\Phi(z_{кр}) + \alpha = 0{,}5$. Откуда $\Phi(z_{кр}) = \frac{1-2\alpha}{2}$.

Для того чтобы найти правую границу правосторонней критической области $z_{кр}$ достаточно найти значение аргумента функции Лапласа, которому соответствует значение функции, равное $\frac{1-2\alpha}{2}$. Тогда правосторонняя критическая область определяется неравенством $Z > z_{кр}$, а область принятия нулевой гипотезы неравенством $Z < z_{кр}$.

Правило 2.

1. Вычислить наблюдаемое значение критерия $Z_{набл}$.
2. По таблице функции Лапласа найти критическую точку из равенства $\Phi(z_{кр}) = \frac{1-2\alpha}{2}$.
3. Если $Z_{набл} < Z_{кр}$ – нет оснований отвергнуть нулевую гипотезу. Если $Z_{набл} > Z_{кр}$ – нулевую гипотезу отвергаем.

Пример.

По двум независимым выборкам объемов $n = 10$ и $m = 10$, извлеченным из нормальных генеральных совокупностей, найдены выборочные средние $\bar{x} = 14{,}3$ и $\bar{y} = 12{,}2$. Генеральные дисперсии известны: $D(X) = 22$, $D(Y) = 18$. При уровне значимости 0,05, проверить нулевую гипотезу H_0: $M(X) = M(Y)$ при конкурирующей гипотезе H_1: $M(X) > M(Y)$.

Решение.

1) Найдем наблюдаемое значение критерия:

$$Z_{набл} = \frac{\bar{x} - \bar{y}}{\sqrt{D(X)/n + D(Y)/n}} = \frac{14{,}3 - 12{,}2}{\sqrt{\frac{22}{10} + \frac{18}{10}}} = 1{,}05$$

2) По условию конкурирующая гипотеза имеет вид H_1: $M(X) > M(Y)$, поэтому критическая область – правосторонняя. Найдем правую критическую точку по равенству

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

По таблице функции Лапласа (Приложение 1) находим $z_{кр} = 1{,}64$.

Так как $Z_{набл} < z_{кр}$ – нет оснований отвергнуть нулевую гипотезу отвергаем, выборочные средние различаются незначимо.

Третий случай.

Нулевая гипотеза H_0: $M(X) = M(Y)$.

Конкурирующая гипотеза H_1: $M(X) < M(Y)$.

В этом случае строят левостороннюю критическую область исходя из требования, чтобы вероятность попадания критерия в эту область, в предположении справедливости нулевой гипотезы, была равна принятому уровню значимости (см. рис. 17): $P\left(Z < z'_{кр}\right) = \alpha$.

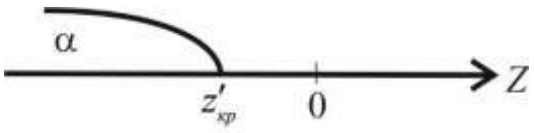


Рис. 17. Левосторонняя критическая область

Так как критерий Z распределен симметрично относительно нуля, искомая критическая точка $z'_{кр}$ симметрична такой точке $z_{кр} > 0$, для которой $P\left(Z > z_{кр}\right) = \alpha$, т. е. $z'_{кр} = -z_{кр}$. Для того чтобы найти точку $z'_{кр}$, достаточно сначала найти "вспомогательную точку" $z_{кр}$, как описано во втором случае, а затем взять найденное значение со знаком минус. Тогда левосторонняя критическая область определяется неравенством $Z < -z_{кр}$, а область принятия нулевой гипотезы – неравенством $Z > -z_{кр}$

Правило 3.

1. Вычислить $Z_{набл}$.
2. По таблице функции Лапласа найти “вспомогательную точку” $z_{кр}$ по равенству $\Phi\left(z_{кр}\right) = \dfrac{1-2\alpha}{2}$, а затем положить $z'_{кр} = -z_{кр}$.
3. Если $Z_{набл} > -z_{кр}$, – нет оснований отвергать нулевую гипотезу. Если $Z_{набл} < -z_{кр}$ – нулевую гипотезу отвергают.

Пример.

По двум независимым выборкам объемов $n = 50$ и $m = 50$, извлеченным из нормальных генеральных совокупностей, найдены выборочные средние $\bar{x} = 142$ и $\bar{y} = 150$. Генеральные дисперсии известны: $D(X) = 28{,}2$

$D(Y) = 22{,}8$. При уровне значимости 0,01, проверить нулевую гипотезу H_0: $M(X) = M(Y)$ при конкурирующей гипотезе H_1: $M(X) < M(Y)$.

Решение.

1) Найдем наблюдаемое значение критерия:

$Z_{набл} = -8$

2) По условию конкурирующая гипотеза имеет вид H_1: $M(X) < M(Y)$, поэтому критическая область – левосторонняя. Найдем "вспомогательную точку" $z_{кр}$ по равенству

$$\Phi(z_{кр}) = \frac{1-2\alpha}{2} = \frac{1-2\cdot 0{,}01}{2} = 0{,}49.$$

По таблице функции Лапласа (Приложение 2) находим $z_{кр} = 2{,}33$. Следовательно $z'_{кр} = -z_{кр} = -2{,}33$

Так как $Z_{набл} < -z_{кр}$ – нулевую гипотезу отвергают отвергаем, выборочные средняя $\bar{x}$ значимо меньше средней $\bar{y}$.

4.5. Сравнение двух средних нормальных генеральных совокупностей, дисперсии которых неизвестны

Пусть генеральные совокупности X и Y распределены *нормально*, причем их дисперсии *неизвестны*. Например, по выборкам малого объема ($n<30$, $m<30$) нельзя получить хорошие оценки генеральных дисперсий. По этой причине метод сравнения средних, изложенный ранее, неприменим. Однако если дополнительно предположить, что неизвестные генеральные дисперсии равны между собой, то можно построить критерий сравнения средних. Например, если сравниваются средние размеры двух партий деталей, изготовленных на одном и том же станке, то естественно допустить, что дисперсии контролируемых размеров одинаковы. Возможен и случай, когда нет оснований считать дисперсии одинаковыми. Тогда перед тем как сравнивать средние, нужно, пользуясь критерием Фишера-Снедекора, проверить гипотезу о равенстве генеральных дисперсий.

Основная задача выглядит следующим образом. В предположении, что генеральные дисперсии одинаковы, требуется проверить нулевую гипотезу H_0: $M(X) = M(Y)$. Т. е. требуется выяснить, значимо или незначимо различаются выборочные средние и, найденные по *независимым малым выборкам* объемов n и m.

В качестве критерия проверки нулевой гипотезы рассмотрим случайную величину

$$T = \frac{\overline{X} - \overline{Y}}{\sqrt{(n-1)s_x^2 + (m-1)s_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}}$$

Величина *T* при справедливости нулевой гипотезы имеет t-распределение Стьюдента с $k = n + m - 2$ степенями свободы.

Первый случай

При нулевой гипотезе H_0: $M(X) = M(Y)$

Конкурирующей будет гипотеза H_1: $M(X) \neq M(Y)$.

В этом случае строят двустороннюю критическую область. При этом, исходят из требования, чтобы вероятность попадания критерия *T* в эту область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости α. Наибольшая мощность критерия достигается тогда, когда "левая" и "правая" критические точки выбраны так, что, $P(T < t_{лев.кр}) = \alpha/2$, $P(T > t_{прав.кр}) = \alpha/2$.

Поскольку случайная величина *T* имеет распределение Стьюдента (симметричное относительно нуля), то и критические точки симметричны относительно нуля. Т.о., если обозначить правую границу *двусторонней* критической области через $t_{двуст.кр}(\alpha; k)$, то левая граница равна $-t_{двуст.кр}(\alpha; k)$

Правило 1.

1) Вычислить наблюдаемое значение критерия:

$$T_{набл} = \frac{\bar{x} - \bar{y}}{\sqrt{(n-1)s_x^2 + (m-1)s_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}}$$

2) По таблице критических точек распределения Стьюдента (Приложение 5) при заданном уровне значимости α и числу степеней свободы $k = n + m - 2$ найти критическую точку $t_{двуст.кр}(\alpha; k)$.

Если окажется, что $|T_{набл}| < t_{двуст.кр}(\alpha; k)$, то отвергать нулевую гипотезу нет оснований. В противном случае ее отвергают.

Пример.

По двум независимым малым выборкам, объемы которых $n = 5$ и $m = 6$, извлеченным из нормальных генеральных совокупностей *X* и *Y*, найдены выборочные средние: $\bar{x} = 3{,}3$, $\bar{y} = 2{,}48$ и исправленные дисперсии:

$s_x^2 = 0{,}25$ и $s_y^2 = 0{,}108$. Требуется при уровне значимости 0,05 проверить нулевую гипотезу H_0: $M(X) = M(Y)$ при конкурирующей гипотезе H_1: $M(X) \neq M(Y)$.

Решение.

1) Исправленные дисперсии различны, поэтому проверим предварительно гипотезу о равенстве генеральных дисперсии, используя критерий Фишера-Снедекора (Приложение 6).

Найдем отношение большей дисперсии к меньшей:

$$F_{набл} = \frac{s_Б^2}{s_M^2} = \frac{0{,}25}{0{,}108} = 2{,}31$$

Дисперсия $s_x^2 = 0{,}25$ значительно больше дисперсии $s_y^2 = 0{,}108$, поэтому в качестве конкурирующей примем гипотезу H_1: $D(X) > D(Y)$.В этом случае критическая область – правосторонняя.

По таблице по уровню значимости α=0,05 и числам степеней свободы $k_1 = n - 1 = 5 - 1 = 4$ и $k_2 = m - 1 = 6 - 1 = 5$ находим критическую точку $F_{Kp}(0{,}05; 4; 4) = 5{,}19$.

Так как $F_{набл} < F_{кр}$ – нет оснований отвергнуть нулевую гипотезу о равенстве генеральных дисперсий. Предположение о равенстве генеральных дисперсий выполняется, поэтому сравним средние.

2) Вычислим наблюдаемое значение критерия Стьюдента:

$$T_{набл} = \frac{\bar{x} - \bar{y}}{\sqrt{(n-1)s_x^2 + (m-1)s_y^2}} \sqrt{\frac{nm(n+m-2)}{n+m}} = 3{,}27$$

По условию, конкурирующая гипотеза имеет вид $M(X) \neq M(Y)$ поэтому критическая область – двусторонняя. По уровню значимости 0,05 и числу степеней свободы $k = 5 + 6 - 2 = 9$ находим по таблице критическую точку $t_{двуст.кр}(0{,}05; 9) = 2{,}26$.

Так как $|T_{набл}| > t_{двуст.кр}(\alpha; k)$ – нулевую гипотезу о равенстве средних отвергаем. Другими словами, выборочные средние различаются значимо.

Второй случай

При нулевой гипотезе H_0: $M(X) = M(Y)$

Конкурирующей будет гипотеза H_1: $M(X) > M(Y)$.

В этом случае строят правостороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия T в эту область

в предположении справедливости нулевой гипотезы была равна принятому уровню значимости α. Наибольшая мощность критерия достигается тогда, когда "левая" и "правая" критические точки выбраны так, что, $P(T > t_{правкр}) = \alpha$.

Критическую точку $t_{правосткр}(\alpha; k)$ находят по таблице (Приложение 5) при уровне значимости α, помещенному в нижней строке таблицы и по числу степеней свободы $k = n + m - 2$.

Если $T_{набл} < t_{правосткр}$, то отвергать нулевую гипотезу нет оснований. В противном случае ее отвергают. Если $T_{набл} > t_{правосткр}$ – нулевую гипотезу отвергают.

Третий случай

При нулевой гипотезе H_0: $M(X) = M(Y)$

Конкурирующей будет гипотеза H_1: $M(X) < M(Y)$.

В этом случае строят левостороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия T в эту область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости:

$$P(T < t_{левост.кр}) = \alpha$$

В силу симметрии распределения Стьюдента относительно нуля $t_{левост.кр} = -t_{правосткр}$. Поэтому сначала находят "вспомогательную" критическую точку $t_{правосткр}$, как описано во втором случае, и полагают $t_{левост.кр} = -t_{правосткр}$.

Если $T_{набл} > -t_{правосткр}$, то отвергать нулевую гипотезу нет оснований. В противном случае ее отвергают. Если $T_{набл} < -t_{правосткр}$ – нулевую гипотезу отвергают.

4.6. Проверка гипотезы о виде распределения

4.6.1. Критерий согласия Пирсона

Особое место занимают гипотезы относительно согласованности выборочного распределения с теоретическим (генеральным) распределением.

Проверка гипотезы о предполагаемом законе неизвестного распределения производится так же, как и проверка гипотезы о параметрах распределения, т. е. при помощи специально подобранной случайной величины – критерия согласия.

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

Критерий согласия позволяет ответить на вопрос о том, является ли различие между выборочными и теоретическим распределениями столь незначительными, что они могут быть приписаны лишь случайным факторам.

Пусть закон распределения генеральной совокупности неизвестен, но есть основания предполагать, что он имеет определенный вид. В частности, если выполняются условия центральной предельной теоремы, есть основания ожидать, что генеральное распределение – нормальное. Если же выборочное среднее и выборочная дисперсия равны, то следует предположить, что генеральная совокупность распределена по закону Пуассона.

Кроме этого, сравнение гистограммы с известными кривыми функций плотностей позволяет также выдвинуть гипотезу о виде распределения генеральной совокупности. Так, исходя из приведенных ниже гистограмм (см. рисунок 18), можно предположить, что исследуемая генеральная совокупность распределена по нормальному (а), показательному (б) и равномерному (в) закону распределения.

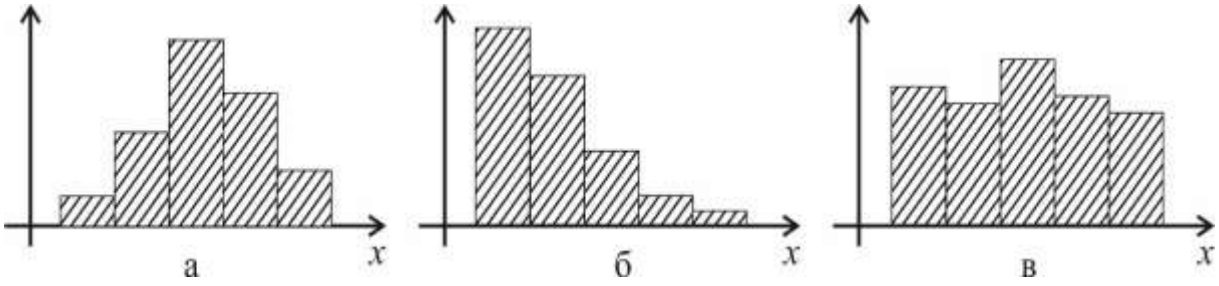


Рис. 18. Сравнение гистограммы с известными кривыми функций плотностей

Эти утверждения носят характер гипотез, а не категорических утверждения, и должны быть подвергнуты статистической проверке.

Для проверки гипотезы закон распределения имеет данный вид (равномерный, нормальный и т. д.), используем *критерий согласия Пирсона*.

С помощью критерия Пирсона можно проверить гипотезу о различных законах распределения генеральной совокупности (равномерном, нормальном, показательном и др.) Для этого в предположении о конкретном виде распределения вычисляются теоретические частоты, и в качестве критерия выбирается случайная величина

$$\chi^2 = \sum \frac{(n_i - n_i')^2}{n_i'}.$$

Эта величина случайная, так как в различных опытах она принимает различные, заранее неизвестные значения. Ясно, что чем меньше различаются эмпирические и теоретические частоты, тем меньше величина критерия и, следовательно, он в известной степени характеризует близость эмпирического и теоретического распределения.

Доказано, что при $n\to\infty$ закон распределения случайной величины независимо от того, какому закону распределения подчинена генеральная совокупность, стремится к закону распределения χ^2 с k степенями свободы. Поэтому случайная величина обозначена через χ^2, а сам критерий называют критерием согласия “хи квадрат”.

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.

Критическая область выбирается правосторонней, и граница ее при заданном уровне значимости α $\chi^2_{кр}(\alpha;k)$ находится по таблице критических точек распределения χ^2.

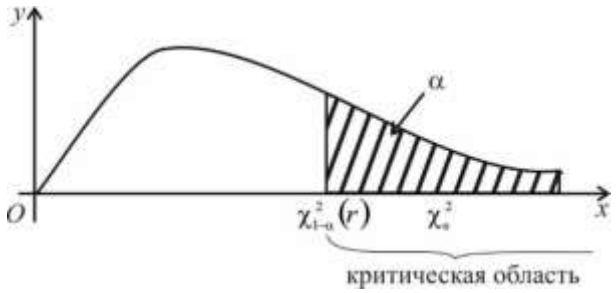


Рис. 19. Критическая область критерия Пирсона

Пусть по выборке объема n получено эмпирическое распределение:

Будем сравнивать эмпирические (наблюдаемые) и теоретические частоты (вычисленные в предположении вида распределения).

Обычно эмпирические и теоретические частоты различаются. Случайно ли расхождение частот? Возможно, что расхождение случайно и объясняется малым числом наблюдений либо способом их группировки, либо другими причинами. Возможно, что расхождение частот неслучайно (значимо) и объясняется тем, что теоретические частоты вычислены исходя из неверной гипотезы о виде распределении генеральной совокупности.

Критерий Пирсона отвечает на поставленный выше вопрос. Как и любой критерий, он не доказывает справедливость гипотезы, а лишь устанавливает на принятом уровне значимости ее согласие или несогласие с данным наблюдением.

4.6.2. Проверка гипотезы о нормальном распределении генеральной совокупности

Допустим, что в предположении нормального распределения генеральной совокупности вычислены теоретические частоты n_i'. При уровне значимости α требуется проверить нулевую гипотезу о нормальном распределении генеральной совокупности.

В качестве критерия проверки нулевой гипотезы примем случайную величину $\chi^2 = \sum \frac{(n_i - n_i')^2}{n_i'}$, где n_i – эмпирические частоты, n_i' – теоретические частоты.

Если предполагаемое распределение нормальное, то оценивают два параметра (математическое ожидание и среднее квадратичное отклонение). Поэтому $r = 2$ и число степеней свободы $k=s-1-r=s-1-2=s-3$.

Алгоритм применения критерия Пирсона

1 Эмпирическое распределение задано в виде последовательности равноотстоящих вариант и соответствующим им частот.

варианты x_i	x_1	x_2	…	x_s
эмпирические частоты n_i	n_1	n_2	…	n_s

Правило 1.

Для того, чтобы при заданном уровне значимости α проверить гипотезу о нормальном распределении генеральной совокупности, надо:

1. Вычислить выборочную среднюю $\bar{x}_в$ и выборочное ско $\sigma_в$.

2. Вычислить теоретические частоты $n_i' = \frac{nh}{\sigma_в}\varphi(u_i)$, где n – объем выборки; h – шаг (разность между двумя соседними вариантами; $u_i = \frac{x_i - \bar{x}_в}{\sigma_в}$; $\varphi(u) = \frac{1}{\sqrt{2\pi}} e^{-\frac{u^2}{2}}$.

3. Сравнить эмпирические и теоретические частоты с помощью критерия Пирсона. Для этого:

а) составляют расчетную таблицу, по которой находят наблюдаемое значение критерия $\chi^2_{набл} = \sum_{i=1}^{k} \frac{(n_i - n_i')^2}{n_i'}$;

б) по таблице критических точек распределения χ^2, по заданному уровню значимости α и числу степеней свободы $k=s-3$ (s - число групп выборки) находят критическую точку $\chi^2_{кр}(\alpha;k)$ правосторонней критической области.

Если $\chi^2_{набл} < \chi^2_{кр}$ – нет оснований отвергнуть гипотезу о нормальном распределении, т. е. эмпирические и теоретические частоты различаются незначимо (случайно). Если $\chi^2_{набл} > \chi^2_{кр}$ – гипотезу отвергают т. е. эмпирические и теоретические частоты различаются значимо.

Замечание: малочисленные частоты ($n_i < 5$) следует объединить; в этом случае и соответствующие им теоретические частоты надо сложить. Если производить объединение частот, то при определении числа степеней свободы в качестве s следует принять число групп, оставшихся после объединения.

Пример.

Используя критерий Пирсона, при уровне значимости 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. Найдем выборочную среднюю $\bar{x}_в = 12{,}63$ и выборочное среднее квадратическое отклонение $\sigma_в = 4{,}695$.

2. Вычислим теоретические частоты, учитывая, что n = 200, h = 2, $\sigma_в = 4{,}695$, по формуле

$$n'_i = \frac{nh}{\sigma_в}\varphi(u_i) = \frac{200 \cdot 2}{4{,}695}\varphi(u_i) = 85{,}2\varphi(u_i)$$

Составим расчетную таблицу:

i	x_i	$u_i = \frac{x_i - \bar{x}_в}{\sigma_в}$	$\varphi(u_i)$	$n'_i = 85{,}2\varphi(u_i)$
1	5	-1,63	0,1057	9,1
2	7	-1,20	0,1942	16,55
3	9	-0,77	0,2966	25,27
4	11	-0,35	0,3752	31,97
5	13	0,08	0,3977	33,88
6	15	0,50	0,3521	30
7	17	0,93	0,2589	22,06
8	19	1,36	0,1582	13,64
9	21	1,78	0,0818	6,97

3. Сравним эмпирические и теоретические частоты

Составим расчетную, из которой найдем наблюдаемое значение критерия

$$\chi^2_{набл} = \sum_{i=1}^{k} \frac{(n_i - n_i')^2}{n_i'}$$

i	n_i	n_i'	$n_i - n_i'$	$(n_i - n_i')^2$	$\frac{(n_i - n_i')^2}{n_i'}$
1	15	9,1	6	36	4
2	26	16,55	9,5	90,25	5,5
3	25	25,27	-0,3	0,09	0,0
4	30	31,97	-2	4	0,13
5	26	33,88	-7,9	62,41	1,8
6	21	30	-9	81	2,7
7	24	22,06	1,94	3,76	0,17
8	20	13,64	6,5	42,25	3,13
9	13	6,97	6	36	5,14
Σ					$\chi^2_{набл} = 22{,}57$

По таблице критических точек распределения χ^2 (Приложение 5) по уровню значимости α= 0,05 и числу степеней свободы k=s-3=9-3 находим критическую точку правосторонней критической области $\chi^2_{кр}(0{,}05;6) = 12{,}6$.

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

2. Эмпирическое распределение задано в виде последовательности интервалов одинаковой длины и соответствующих им частот.

Пусть эмпирическое распределение задано в виде последовательности интервалов $(x_i; x_{i+1})$ и соответствующих им частот n_i.

$(x_i; x_{i+1})$	$(x_1; x_2)$	$(x_2; x_3)$	…	$(x_s; x_{s+1})$
n_i	n_1	n_2	…	n_s

Требуется, используя критерий Пирсона, проверить гипотезу о том, что генеральная совокупность X распределена нормально.

Правило 2

Для того чтобы при уровне значимости αпроверить гипотезу о нормальном распределении генеральной совокупности надо:

1) Вычислить выборочную среднюю $\bar{x}_в$ и выборочное ско $\sigma_в$ причем в качестве вариант x_i^* принимают среднее арифметическое концов интервала:

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

2) Пронормировать X, т. е. перейти к случайной величине $Z = \dfrac{X - \bar{x}_в}{\sigma_в}$

и вычислить концы интервалов $z_i = \dfrac{x_i - \bar{x}_в}{\sigma_в}$, $z_i = \dfrac{x_{i+1} - \bar{x}_в}{\sigma_в}$.

3) Вычислить теоретические частоты $n_i' = n \cdot P_i$ где n – объем выборки (сумма всех частот); $P_i = \Phi(z_{i+1}) - \Phi(z_i)$ – вероятности попадания X в интервалы (x_i, x_{i+1}); Ф(z) – функция Лапласа.

4) Сравнить эмпирические и теоретические частоты с помощью критерия Пирсона. Для этого:

а) составляют расчетную таблицу, по которой находят наблюдаемое значение критерия Пирсона $\chi^2_{набл} = \sum_{i=1}^{k} \dfrac{(n_i - n_i')^2}{n_i'}$.

б) по таблице критических точек распределения χ^2 по заданному уровню значимости αи числу степеней свободы k=s-3 (s – число интервалов выборки) находят критическую точку правосторонней критической области $\chi^2_{кр}(\alpha; k)$.

Если $\chi^2_{набл} < \chi^2_{кр}$ – нет оснований отвергнуть гипотезу о нормальном распределении генеральной совокупности. Если $\chi^2_{набл} > \chi^2_{кр}$ – гипотезу отвергают.

Замечание. Интервалы, содержащие малочисленные эмпирические частоты n_i < 5), следует объединить, а частоты этих интервалов сложить. Если производилось объединение интервалов, то при определении числа степеней следует в качестве s принять число интервалов, оставшихся после объединения интервалов.

Пример

Используя критерий Пирсона, при уровне значимости 0,05 проверить, согласуется ли гипотеза о нормальном распределении генеральной совокупности X с эмпирическим распределением выборки объема n = 100.

	1	2	3	4	5	6	7
$(x_i; x_{i+1})$	3-8	8-13	13-18	18-23	23-28	28-33	33-38
n_i	6	8	15	40	16	8	7

Решение.

1. Вычислим выборочную среднюю и выборочное среднее квадратическое отклонение. Для этого перейдем от заданного интервального распределения к распределению равноотстоящих вариант, приняв в качестве варианты x_i^* среднее арифметическое концов интервала. В итоге получим распределение:

x_i^*	5,5	10,5	15,5	20,5	25,5	30,5	35,5
n_i	6	8	15	40	16	8	7

$\bar{x}_в = 20{,}7;\ \sigma_в = 7{,}28.$

2. Найдем интервалы $(z_i;z_{i+1})$

i	Границы интервала		$x_i - \bar{x}_в$	$x_{i+1} - \bar{x}_в$	Границы интервала	
	x_i	x_{i+1}			$z_i = \frac{x_i - \bar{x}_в}{\sigma_в}$	$z_i = \frac{x_{i+1} - \bar{x}_в}{\sigma_в}$
1	3	8	–	–12,7	–∞	–1,74
2	8	13	–12,7	–7,7	–1,74	–1,06
3	13	18	–7,7	–2,7	–1,06	–0,37
4	18	23	–2,7	2,3	–0,37	0,32
5	23	28	2,3	7,3	0,32	1,00
6	28	33	7,3	12,3	1,00	1,69
7	33	38	12,3	–	1,69	+∞

3. Найдем теоретические вероятности P_i, и теоретические частоты $n_i' = n \cdot P_i = 100P_i$ Для этого составим расчетную таблицу. z_i

i	Границы интервала		$\Phi(z_i)$	$\Phi(z_{i+1})$	$P_i = \Phi(z_{i+1}) - \Phi(z_i)$	$n_i' = 100P_i$
	z_i	z_{i+1}				
1	–∞	–1,74	–0,5000	–0,4591	0,0409	4,09
2	–1,74	–1,06	–0,4591	–0,3554	0,1037	10,37
3	–1,06	–0,37	–0,3554	–0,1443	0,2111	21,11
4	–0,37	0,32	–0,1443	0,1255	0,2698	26,98
5	0,32	1,00	0,1255	0,3413	0,2158	21,58
6	1,00	1,69	0,3413	0,4545	0,1132	11,32
7	1,69	+∞	0,4545	0,5000	0,0455	4,55
Σ					1	100

4. Сравним эмпирические и теоретические частоты, используя критерий Пирсона.

а) вычислим наблюдаемое значение критерия Пирсона. Для этого составим расчетную табл. Столбцы 7 и 8 служат для контроля вычислений по формуле $\chi^2_{набл} = \sum_{i=1}^{k} \frac{(n_i)^2}{n'_i} - n$

	n_i	n'_i	$n_i - n'_i$	$(n_i - n'_i)^2$	$\frac{(n_i - n'_i)^2}{n'_i}$	n_i^2	$\frac{n_i^2}{n'_i}$
1	6	4,09	1,91	3,6481	0,8920	36	8,8019
2	8	10,37	–2,37	5,6169	0.5416	64	6,1716
3	15	21,11	–6,11	37,3121	1,7684	225	10,6584
4	40	26,98	13,02	169,5204	6,2833	1600	59,3052
5	16	21,58	–5,58	31,1364	1,4428	256	11,8628
6	8	11,32	–3,32	11,0224	0,9737	64	5,6537
7	7	4,55	2,45	6,0025	1,3192	49	10.7692
Σ	100	100			$\chi^2_{набл} = 13,22$		113,22

Контроль: $\chi^2_{набл} = \sum_{i=1}^{k} \frac{(n_i)^2}{n'_i} - n = 113,22 - 100 = 13,22 = \chi^2_{набл}$

Вычисления произведены правильно.

б) По таблице критических точек распределения χ^2 по уровню значимости $\alpha = 0,05$ и числу степеней свободы $k = s - 3 = 7 - 3 = 4$ находим критическую точку правосторонней критической области $\chi^2_{кр}(0,05;4) = 9,5$.

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

4.6.3. Проверка гипотезы о показательном распределении генеральной совокупности

Задано эмпирическое распределение непрерывной случайной величины *X* в виде последовательности интервалов $(x_i; x_{i+1})$ и соответствующих им частот n_i, причем $\sum n_i = n$ (объем выборки). Объем выборки должен быть достаточно велик ($n \geq 50$). Требуется, используя критерий Пирсона, проверить гипотезу о том, что случайная величина *X* имеет показательное распределение.

Правило. Для того чтобы при уровне значимости α проверить гипотезу о том, что непрерывная случайная величина распределена по показательному закону, надо:

1. Найти по заданному эмпирическому распределению выборочную среднюю $\overline{x}_в$. Для этого, приняв в качестве "представителя" *i*-го интервала его середину $x_i^* = \frac{x_i + x_{i+1}}{2}$, составляют последовательность равноотстоящих вариант и соответствующих им частот.

2. Принять в качестве оценки параметра λ показательного распределения величину, обратную выборочной средней:

$$\lambda^* = \frac{1}{\overline{x}_в}$$

3. Найти вероятности попадания X в частичные интервалы $(x_i; x_{i+1})$ по формуле $P_i = P(x_i < X < x_{i+1}) = e^{-\lambda x_i} - e^{-\lambda x_{i+1}}$

4. Вычислить теоретические частоты: $n_i' = n \cdot P_i$.

5. Сравнить эмпирические и теоретические частоты с помощью критерия Пирсона, приняв число степеней свободы $k = s - 2$, где s – число первоначальных интервалов выборки; если же было произведено объединение малочисленных частот, следовательно, и самих интервалов, то s – число интервалов, оставшихся после объединения.

Пример.

В результате испытания 200 элементов на длительность работы получено эмпирическое распределение (в первом столбце указаны интервалы времени в часах, во втором столбце – частоты, т. е. количество элементов, проработавших время в пределах соответствующего интервала).

$(x_i; x_{i+1})$	0–5	5–10	10–15	15–20	20–25	25–30
n_i	133	45	15	4	2	1

Требуется, при уровне значимости 0,05, проверить гипотезу о том, что время работы элементов распределено по показательному закону.

Решение.

1) Найдем среднее время работы всех элементов. Перейдем к распределению середин интервалов.

x^*	2,5	7,5	12,5	17,5	22,5	27,5
n_i	133	45	15	4	2	1

$$\bar{x}_в = \frac{1}{200}(2{,}5\cdot 133 + 7{,}8\cdot 45 + 12{,}5\cdot 15 + 17{,}5\cdot 4 + 22{,}5\cdot 2 + 27{,}5\cdot 1) = 5$$

2) Найдем оценку параметра предполагаемого показательного распределения:

$$\lambda^* = \frac{1}{5} = 0{,}2.$$

Таким образом, плотность предполагаемого показательного распределения имеет вид

$$f(x) = 0{,}2e^{-0{,}2x} \ (x > 0).$$

3) Найдем вероятности попадания X в каждый из интервалов по формуле

$$P_i = P(x_i < X < x_{i+1}) = e^{-\lambda x_i} - e^{-\lambda x_{i+1}}.$$

Например, для первого интервала

$$P_1 = P(0 < X < 5) = e^{-0{,}2\cdot 0} - e^{-0{,}2\cdot 5} = 1 - e^{-1} = 1 - 0{,}3679 = 0{,}6321.$$

Аналогично вычислим вероятности попадания X в остальные интервалы. Результаты заносим в таблицу.

4) Найдем теоретические частоты: $n'_i = n \cdot P_i = 200 \cdot P_i$. Результаты заносим в таблицу.

5) Сравним эмпирические и теоретические частоты с помощью критерия Пирсона. Для этого составим расчетную таблицу, причем объединим малочисленные частоты (4+2+1=7) и соответствующие им теоретические частоты (6,30 + 2,32 + 0,84 = 9,46). Результаты заносим в таблицу.

i	n_i	n'_i	$n_i - n'_i$	$(n_i - n'_i)^2$	$\frac{(n_i - n'_i)^2}{n'_i}$
1	133	126,42	6,58	43,2964	0,3425
2	45	46,52	–1,52	2,3104	0,0497
3	15	17,10	–2,10	4,4100	0,2579
4	7	9,46	–2,46	6,0516	0,6397
Σ	n = 200				$\chi^2_{набл} = 1{,}29$

Замечание.

Для упрощения вычислений в случае объединения малочисленных частот целесообразно объединить и сами интервалы, которым принадлежат малочисленные частоты, в один интервал. Так, в рассматриваемой задаче, объединив последние три интервала, получим один интервал (15, 30). В этом случае теоретическая частота

$$n'_4 = n \cdot P_4(15 < X < 30) = 200 \cdot 0{,}0473 = 9{,}6$$

Находим $\chi^2_{набл} = 1{,}29$. По таблице критических точек распределения χ^2 (см. приложение 5), по уровню значимости α = 0,05 и числу степеней свободы $k = s - 2 = 4 - 2 = 2$ находим критическую точку правосторонней критической области $\chi^2_{кр}(0{,}05;2) = 6$.

Так как $\chi^2_{набл} < \chi^2_{кр}$ – нет оснований отвергнуть гипотезу о распределении *X* по показательному закону. Другими словами, данные наблюдений согласуются с этой гипотезой.

4.6.4. Проверка гипотезы о равномерном распределении генеральной совокупности

Задано эмпирическое распределение непрерывной случайной величины *X* в виде последовательности интервалов $(x_i; x_{i+1})$ и соответствующих им частот n_i, причем $\sum n_i = n$ (объем выборки). Объем выборки должен быть достаточно велик ($n \geq 50$). Требуется, используя критерий Пирсона, проверить гипотезу о том, что случайная величина *X* имеет показательное распределение.

Правило. Для того чтобы проверить гипотезу о равномерном распределении X, т. е. по закону

$$f(x) = \begin{cases} \dfrac{1}{b-a} \text{ в интервале } (a;b), \\ 0 \text{ вне интервале } (a;b) \end{cases}$$

надо:

1. Оценить параметры *a* и *b* – концы интервала, в котором наблюдались возможные значения *X*, по формулам (через a^* и b^** обозначены оценки параметров):

$$a^* = \bar{x}_в - \sqrt{3}\sigma_в, \; b^* = \bar{x}_в + \sqrt{3}\sigma_в$$

2. Найти плотность вероятности предполагаемого распределения

$$f(x) = \frac{1}{b^* - a^*}$$

3. Найти теоретические частоты:

$$n'_1 = nP_1 = n\left(f(x) \cdot \left(x_1 - a^*\right)\right) = n \cdot \frac{1}{b^* - a^*} \cdot \left(x_1 - a^*\right);$$

$$n'_2 = n'_3 = \ldots = n'_{s-1} = n \cdot \frac{1}{b^* - a^*} \cdot \left(x_i - x_{i-1}\right), (1 = 1,2,\ldots,s-1)$$

$$n'_s = n \cdot \frac{1}{b^* - a^*} \cdot \left(b^* - x_{s-1}\right)$$

4. Сравнить эмпирические и теоретические частоты с помощью критерия Пирсона, приняв число степеней свободы $k = s - 3$, где s – число интервалов, на которые разбита выборка.

Пример.

Произведено n = 200 испытаний, в результате каждого из которых событие A появлялось в различные моменты времени. В итоге было получено эмпирическое распределение, приведенное в таблице (в первом столбце указаны интервалы времени в минутах, во втором столбце – соответствующие частоты, т. е. число появлений события A в интервале). Требуется при уровне значимости 0,05 проверить гипотезу о том, что время появления событий распределено равномерно.

$(x_i; x_{i+1})$	n_i	$(x_i; x_{i+1})$	n_i
2 – 4	21	12 – 14	14
4 – 6	16	14 – 16	21
6 – 8	15	16 – 18	22
8 – 10	26	18 – 20	18
10 – 12	22	20 – 22	25

Решение.

1) Найдем оценки параметров a и b равномерною распределения по формулам:

$$a^* = \bar{x}_в - \sqrt{3}\sigma_в,\ b^* = \bar{x}_в + \sqrt{3}\sigma_в$$

Для вычисления выборочной средней $\bar{x}_в$ и выборочного среднего квадратического отклонения $\sigma_в$ перейдем к распределению середин интервалов.

x^*	3	5	7	9	11	13	15	17	19	21
n_i	21	16	15	26	22	14	21	22	18	25

$\bar{x}_в = 12{,}31;\ \sigma_в = 5{,}81.$

Следовательно,

$a^* = 12{,}31 - \sqrt{3} \cdot 5{,}81 = 2{,}26;\ b^* = 12{,}31 + \sqrt{3} \cdot 5{,}81 = 22{,}36$

a* - 12,31 –1,73 -5,81 =2,26, 6* = 12,31 1,73 -5,81 =22,36.

2) Найдем плотность предполагаемого равномерного распределения:

$$f(x) = \frac{1}{b^* - a^*} = \frac{1}{22{,}36 - 2{,}26} = 0{,}05.$$

3) Найдем теоретические частоты:

$$n_1' = 200 \cdot 0{,}05 \cdot (4 - 2{,}26);$$

$$n_2' = 200 \cdot 0{,}05 \cdot (x_2 - x_1) = 10 \cdot (6 - 4) = 20.$$

Длины третьего-девятого интервалов равны длине второго интервала, поэтому теоретические частоты, соответствующие этим интервалам и теоретическая частота второго интервала одинаковы, т. е.

$$n'_2 = n'_3 = n'_4 = n'_4 = n'_5 = n'_6 = n'_7 = n'_8 = n'_9 = 20$$

$$n'_{10} = 200 \cdot \frac{1}{22{,}36 - 2{,}26} \cdot (22{,}36 - 20) = 23{,}6$$

4) Сравним эмпирические и теоретические частоты, используя критерий Пирсона, приняв число степеней свободы $k = s - 3 = 10 - 3 = 7$. Для этого составим расчетную таблицу.

i	n_i	n'_i	$n_i - n'_i$	$(n_i - n'_i)^2$	$\frac{(n_i - n'_i)^2}{n'_i}$
1	21	17,3	3,7	13,69	0,79
2	16	20	–4	16	0,80
3	15	20	–5	25	1,25
4	26	20	6	36	1,80
5	22	20	2	4	0,20
6	14	20	–6	36	1,80
7	21	20	1	1	0,05
8	22	20	2	4	0,20
9	18	20	–2	4	0,20
10	25	23,6	1,4	1,96	0,08
Σ	n = 200				$\chi^2_{набл} = 7{,}17$

Из расчетной таблицы получаем $\chi^2_{набл} = 7{,}17$.

Найдем по таблице критических точек распределения χ^2 (см. приложение 5) по уровню значимости α = 0,05 и числу степеней свободы $k = 7$ критическую точку правосторонней критической области $\chi^2_{кр}(0{,}05;7) = 14{,}1$.

Так как $\chi^2_{набл} < \chi^2_{кр}$ – нет оснований отвергнуть гипотезу о распределении X по показательному закону. Другими словами, данные наблюдений согласуются с этой гипотезой.

4.7. Проверка значимости коэффициента корреляции

Так как выборочный коэффициент r вычисляется по выборочным данным, то он является случайной величиной. Если $r \neq 0$, то возникает вопрос: объясняется ли это действительно существующей линейной связью между X и Y или вызвано случайными факторами?

Пусть двумерная генеральная совокупность (*X, Y*) распределена нормально. Из этой совокупности извлечена выборка объема *n* и по ней найден выборочный коэффициент корреляции $r_в \neq 0$. Требуется проверить нулевую гипотезу $H_0 : r_Г = 0$ о равенстве нулю генерального коэффициента корреляции.

Если нулевая гипотеза принимается, то это означает, что *X* и *Y* некоррелированы; в противном случае – коррелированы.

Правило. Для проверки гипотезы используется *t*–критерий Стьюдента. Для того чтобы при уровне значимости α проверить нулевую гипотезу о равенстве нулю генерального коэффициента корреляции нормальной двумерной случайной величины при конкурирующей гипотезе $H_1 : r_Г \neq 0$, надо:

1) Вычислить наблюдаемое значение критерия

$$T_{набл} = r_в \sqrt{\frac{n-2}{1-r_в^2}}$$

где $r_в$ – выборочный коэффициент корреляции;

n – объем выборки.

2) По таблице критических точек распределения Стьюдента, по заданному уровню значимости α и числу степеней свободы $k = n-2$ найти критическую точку $t_{кр}(\alpha; k)$ двусторонней критической области.

3) Вычисленное наблюдаемое значение $T_{набл}$ сравнивается с найденным по таблице критическим значением $t_{кр}(\alpha; k)$. Если $|T_{набл}| < t_{кр}$ – нет оснований отвергнуть нулевую гипотезу. Если $|T_{набл}| > t_{кр}$ – нулевую гипотезу отвергают.

Пример.

По выборке объема *n* =100, извлеченной из двумерной нормальной генеральной совокупности (*X, Y*), найден выборочный коэффициент корреляции $r_в = 0{,}2$. Требуется при уровне значимости 0,05 проверить нулевую гипотезу о равенстве нулю генерального коэффициента корреляции $H_0 : r_Г = 0$ при конкурирующей гипотезе $H_1 : r_Г \neq 0$

Решение.

1) Найдем наблюдаемое (эмпирическое) значение критерия:

$$T_{набл} = r_в \sqrt{\frac{n-2}{1-r_в^2}} = 0{,}2\sqrt{\frac{100-2}{1-0{,}2^2}} = 2{,}02$$

2) По условию, конкурирующая гипотеза имеет вид $H_1 : r_Г \neq 0$ поэтому критическая область – двусторонняя.

3) По таблице критических точек распределения Стьюдента (см. приложение 6), по уровню значимости $\alpha = 0{,}05$, помещенному в верхней строке таблицы, и числу степеней свободы $k = n - 2 = 98$ находим критическую точку двусторонней критической области $t_{кр}(0{,}05;98) = 1{,}99$.

Так как $|T_{набл}| > t_{кр}$ – отвергаем нулевую гипотезу о равенстве нулю генерального коэффициента корреляции. Другими словами, коэффициент корреляции значимо отличается от нуля; следовательно, *X* и *Y* коррелированы.

Вопросы для самопроверки

1) Какая гипотеза называется статистической? Приведите пример.

2) Какая статистическая гипотеза называется нулевой? Альтернативной? Приведите примеры.

3) Что такое уровень значимости? Как он связан с доверительной вероятностью?

4) Что такое критическая область критерия?

5) Поясните смысл ошибок первого и второго рода, возникающих при проверке гипотез.

6) Как связаны вид альтернативной гипотезы и тип критической области?

Задачи для самостоятельного решения

1. Имеются независимые выборки значений нормально распределенных случайных величин

X: 2, 2, 3, 3, 4, 4, 4, 5, 5, 6

Y: 3, 3, 3, 4, 4, 4, 4, 5, 5, 5, 5, 5, 7, 8, 9.

Требуется проверить для уровня значимости $\alpha = 0{,}1$ при условии равенства генеральных дисперсий нулевую гипотезу H_0: $M(X) = M(Y)$ при конкурирующей гипотезе H_1: $M(X) \neq M(Y)$.

2. Проведено исследование розничного товарооборота продовольственных магазинов в двух районах области (по 50 магазинов в каждом). Априори известны средние значения розничного товарооборота – 78,9

и 78,68 тыс. руб. Полученные в результате оценки среднеквадратичных отклонений в первом и втором районах области соответственно равны 7,22 и 7,79 тыс. руб. Можно ли считать, что разброс розничного товарооборота магазинов в районах неодинаков при уровне значимости 0,05? Можно ли сделать вывод о разной покупательной способности населения районов?

3. В банке в течение двух дней проводилось исследование времени обслуживания клиентов. Данные представлены в табл. Обозначим время обслуживания клиентов в первый день X, а во второй – Y.

Номер интервала	Время обслуживания (мин)	n_i (1-й день)	m_i (2-й день)
1.	10-12	2	2
2.	12-14	4	4
3.	14-16	8	9
4.	16-18	12	13
5.	18-20	16	16
6.	20-22	10	8
7.	22-24	3	3

Можно ли считать одинаковыми отклонения от среднего времени обслуживания клиентов банка в 1-й и во 2-й дни при $\alpha = 0{,}1$? Можно ли считать, что среднее время обслуживания равно в первый и второй день на том же уровне значимости.

4. Для выборки, интервальный статистический ряд которой имеет вид

Номер интервала	Границы интервала	Эмпирические частоты
1	2–5	6
2	5–8	8
3	8–11	15
4	11–14	22
5	14–17	14
6	17–20	5

Проверить при уровне значимости 0,05 гипотезу:

а) о показательном; б) равномерном; в) нормальном законе распределения генеральной совокупности с помощью критерия Пирсона.

5. Установить закон распределения признака X – затраты времени на обработку одной детали. Дано распределение 100 рабочих по затратам времени на обработку одной детали (мин):

x_{i-1}–x_i	22–24	24–26	26–28	28–30	30–32	32–34
n_i	2	12	34	40	10	2

6. В таблице представлены результаты испытаний двух случайных величин X и Y. Требуется проверить значимость выборочного коэффициента корреляции.

x_i	4,08	6,91	7,42	3,58	5,16	5,19	4,1	5,37	5,02	6,19
y_i	2,14	3	1,73	4,24	3,27	2,83	4,22	4,4	2,19	3,2

СПИСОК ЛИТЕРАТУРЫ

1. Балдин, К. В. Основы теории вероятностей и математической статистики : учебник / К. В. Балдин, В. Н. Башлыков, А. В. Рокосуев ; ред. К. В. Балдина. – Москва : Издательство "Флинта", 2010. – 245 с.

2. Гмурман, В. Е. Руководство к решению задач по теории вероятностей и математической статистике : учебное пособие для бакалавриата и специалитета / В. Е. Гмурман. – 11-е изд., перераб. и доп. – Москва : Изд-во Юрайт, 2019. – 406 с.

3. Гмурман, В. Е. Теория вероятностей и математическая статистика : учебник для прикладного бакалавриата / В. Е. Гмурман. – 12-е изд. – Москва : Издательство Юрайт, 2019. – 479 с.

4. Голодная Н. Ю., Одияко Н. Н. математическая статистика: Теория корреляции в экономических расчетах. Ч. 2.: Учебное пособие. – Владивосток: Изд-во ВГУЭС, 2006. – 80 с.

5. Гусева, Е. Н. Теория вероятностей и математическая статистика : учебное пособие / Е. Н. Гусева. – 6-е изд., стереотип. – Москва : Издательство "Флинта", 2016. – 220 с.

6. Дегтярь, Л. А. Теория вероятностей и математическая статистика: учебное пособие для самостоятельной работы студентов / Л. А. Дегтярь, А. Г. Мордкович. – пос. Персиановский: Донской ГАУ, 2013. – 108с.

7. Корнеева Е. Н. Математика. Математическая статистика: учебно-методическое пособие/ Е. Н. Корнеева. – Орел: ОрелГТУ, 2010. – 26 с.

8. Кремер Н. Ш. Теория вероятностей и математическая статистика : учебник и практикум для академического бакалавриата / Н. Ш. Кремер. – 5-е изд., перераб. и доп. – Москва : Издательство Юрайт, 2019. – 538 с.

9. Математическая статистика : учеб.-метод. пособие / авт. сост. : С. Е. Демин; М-во образования и науки РФ; ФГАОУ ВО "УрФУ им. первого Президета Росии Б. Н. Ельцина", Нижнетагил. технол. ин-т (фил.). – Нижний Тагил : НТИ (филиал) УрФУ

10. Трофимова Е. А. Теория вероятностей и математическая статистика: учеб. пособие / Е. А. Трофимова, Н. В. Кисляк, Д. В. Гилев; (под общей ред. Е. А. Трофимой); М-во образования и науки Рос. Федерации, Урал. федер. ун-т. – Екатеренбург: Изд-во Урал. Ун-та, 2018. – 160 с.

11. Фастовец Н. О., Попов М. А. Математичесакая статистика: примеры, задачи и типовые задания. Учеб. пособие для нефтегазового образования.

Приложение 1

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

	0	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
0,0	0,3989	0,3989	0,3989	0,3988	0,3986	0,3984	0,3982	0,3980	0,3977	0,3973
0,1	0,3970	0,3965	0,3961	0,3956	0,3951	0,3945	0,3939	0,3932	0,3925	0,3918
0,2	0,3910	0,3902	0,3894	0,3885	0,3876	0,3867	0,3857	0,3847	0,3836	0,3825
0,3	0,3814	0,3802	0,3790	0,3778	0,3765	0,3752	0,3739	0,3726	0,3712	0,3698
0,4	0,3683	0,3668	0,3652	0,3637	0,3621	0,3605	0,3589	0,3572	0,3555	0,3538
0,5	0,3521	0,3503	0,3485	0,3467	0,3448	0,3429	0,3410	0,3391	0,3372	0,3352
0,6	0,3332	0,3312	0,3292	0,3271	0,3251	0,3230	0,3209	0,3187	0,3166	0,3144
0,7	0,3123	0,3101	0,3079	0,3056	0,3034	0,3011	0,2989	0,2966	0,2943	0,2920
0,8	0,2897	0,2874	0,2850	0,2827	0,2803	0,2780	0,2756	0,2732	0,2709	0,2685
0,9	0,2661	0,2637	0,2613	0,2589	0,2565	0,2541	0,2516	0,2492	0,2468	0,2444
1,0	0,2420	0,2396	0,2371	0,2347	0,2323	0,2299	0,2275	0,2251	0,2227	0,2203
1,1	0,2179	0,2155	0,2131	0,2107	0,2083	0,2059	0,2036	0,2012	0,1989	0,1965
1,2	0,1942	0,1919	0,1895	0,1872	0,1849	0,1826	0,1804	0,1781	0,1758	0,1736
1,3	0,1714	0,1691	0,1669	0,1647	0,1626	0,1604	0,1582	0,1561	0,1539	0,1518
1,4	0,1497	0,1476	0,1456	0,1435	0,1415	0,1394	0,1374	0,1354	0,1334	0,1315
1,5	0,1295	0,1276	0,1257	0,1238	0,1219	0,1200	0,1182	0,1163	0,1145	0,1127
1,6	0,1109	0,1092	0,1074	0,1057	0,1040	0,1023	0,1006	0,0989	0,0973	0,0957
1,7	0,0940	0,0925	0,0909	0,0893	0,0878	0,0863	0,0848	0,0833	0,0818	0,0804
1,8	0,0790	0,0775	0,0761	0,0748	0,0734	0,0721	0,0707	0,0694	0,0681	0,0669
1,9	0,0656	0,0644	0,0632	0,0620	0,0608	0,0596	0,0584	0,0573	0,0562	0,0551
2,0	0,0540	0,0529	0,0519	0,0508	0,0498	0,0488	0,0478	0,0468	0,0459	0,0449
2,1	0,0440	0,0431	0,0422	0,0413	0,0404	0,0395	0,0387	0,0379	0,0371	0,0363
2,2	0,0353	0,0347	0,0339	0,0332	0,0325	0,0317	0,0310	0,0303	0,0297	0,0290
2,3	0,0283	0,0277	0,0270	0,0264	0,0258	0,0252	0,0246	0,0241	0,0235	0,0229
2,4	0,0224	0,0219	0,0213	0,0208	0,0203	0,0198	0,0194	0,0189	0,0184	0,0180
2,5	0,0175	0,0171	0,0167	0,0163	0,0158	0,0154	0,0151	0,0147	0,0143	0,0139
2,6	0,0136	0,0132	0,0129	0,0126	0,0122	0,0119	0,0116	0,0113	0,0110	0,0107
2,7	0,0104	0,0101	0,0099	0,0096	0,0093	0,0091	0,0088	0,0086	0,0084	0,0081
2,8	0,0079	0,0077	0,0075	0,0073	0,0071	0,0069	0,0067	0,0065	0,0063	0,0061
2,9	0,0060	0,0058	0,0056	0,0055	0,0053	0,0051	0,0050	0,0048	0,0047	0,0046
3,0	0,0044	0,0043	0,0042	0,0040	0,0039	0,0038	0,0037	0,0036	0,0035	0,0034
3,1	0,0033	0,0032	0,0031	0,0030	0,0029	0,0028	0,0027	0,0026	0,0025	0,0025
3,2	0,0024	0,0023	0,0022	0,0022	0,0021	0,0020	0,0020	0,0019	0,0018	0,0018
3,3	0,0017	0,0017	0,0016	0,0016	0,0015	0,0015	0,0014	0,0014	0,0013	0,0013
3,4	0,0012	0,0012	0,0012	0,0011	0,0011	0,0010	0,0010	0,0010	0,0009	0,0009
3,5	0,0009	0,0008	0,0008	0,0008	0,0008	0,0007	0,0007	0,0007	0,0007	0,0006
3,6	0,0006	0,0006	0,0006	0,0005	0,0005	0,0005	0,0005	0,0005	0,0005	0,0004
3,7	0,0004	0,0004	0,0004	0,0004	0,0004	0,0004	0,0003	0,0003	0,0003	0,0003
3,8	0,0003	0,0003	0,0003	0,0003	0,0003	0,0002	0,0002	0,0002	0,0002	0,0002
3,9	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0002	0,0001

Приложение 2

Таблица значений интегральной функции Лапласа

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_0^x e^{-z^2/2} dz$$

x	*Φ(x)*	*x*	*Φ(x)*	*x*	*Φ(x)*	*x*	*Φ(x)*	*x*	*Φ(x)*	*x*	*Φ(x)*
0,00	0,0000	**0,46**	0,1772	**0,92**	0,3212	**1,38**	0,4162	**1,84**	0,4671	**2,60**	0,4953
0,01	0,0040	**0,47**	0,1808	**0,93**	0,3238	**1,39**	0,4177	**1,85**	0,4678	**2,62**	0,4956
0,02	0,0080	**0,48**	0,1844	**0,94**	0,3264	**1,40**	0,4192	**1,86**	0,4686	**2,64**	0,4959
0,03	0,0120	**0,49**	0,1879	**0,95**	0,3289	**1,41**	0,4207	**1,87**	0,4693	**2,66**	0,4961
0,04	0,0160	**0,50**	0,1915	**0,96**	0,3315	**1,42**	0,4222	**1,88**	0,4699	**2,68**	0,4963
0,05	0,0199	**0,51**	0,1950	**0,97**	0,3340	**1,43**	0,4236	**1,89**	0,4706	**2,70**	0,4965
0,06	0,0239	**0,52**	0,1985	**0,98**	0,3365	**1,44**	0,4251	**1,90**	0,4713	**2,72**	0,4967
0,07	0,0279	**0,53**	0,2019	**0,99**	0,3389	**1,45**	0,4265	**1,91**	0,4719	**2,74**	0,4969
0,08	0,0319	**0,54**	0,2054	**1,00**	0,3413	**1,46**	0,4279	**1,92**	0,4726	**2,76**	0,4971
0,09	0,0359	**0,55**	0,2088	**1,01**	0,3438	**1,47**	0,4292	**1,93**	0,4732	**2,78**	0,4973
0,10	0,0398	**0,56**	0,2123	**1,02**	0,3461	**1,48**	0,4306	**1,94**	0,4738	**2,80**	0,4974
0,11	0,0438	**0,57**	0,2157	**1,03**	0,3485	**1,49**	0,4319	**1,95**	0,4744	**2,82**	0,4976
0,12	0,0478	**0,58**	0,2190	**1,04**	0,3508	**1,50**	0,4332	**1,96**	0,4750	**2,84**	0,4977
0,13	0,0517	**0,59**	0,2224	**1,05**	0,3531	**1,51**	0,4345	**1,97**	0,4756	**2,86**	0,4979
0,14	0,0557	**0,60**	0,2257	**1,06**	0,3554	**1,52**	0,4357	**1,98**	0,4761	**2,88**	0,4980
0,15	0,0596	**0,61**	0,2291	**1,07**	0,3577	**1,53**	0,4370	**1,99**	0,4767	**2,90**	0,4981
0,16	0,0636	**0,62**	0,2324	**1,08**	0,3599	**1,54**	0,4382	**2,00**	0,4772	**2,92**	0,4982
0,17	0,0675	**0,63**	0,2357	**1,09**	0,3621	**1,55**	0,4394	**2,02**	0,4783	**3,00**	0,49865
0,18	0,0714	**0,64**	0,2389	**1,10**	0,3643	**1,56**	0,4406	**2,04**	0,4793	**3,20**	0,49931
0,19	0,0753	**0,65**	0,2422	**1,11**	0,3665	**1,57**	0,4418	**2,06**	0,4803	**3,40**	0,49966
0,20	0,0793	**0,66**	0,2454	**1,12**	0,3686	**1,58**	0,4429	**2,08**	0,4812	**3,60**	0,499841
0,21	0,0832	**0,67**	0,2486	**1,13**	0,3708	**1,59**	0,4441	**2,10**	0,4821	**3,80**	0,499928
0,22	0,0871	**0,68**	0,2517	**1,14**	0,3729	**1,60**	0,4452	**2,12**	0,4830	**4,00**	0,499968
0,23	0,0910	**0,69**	0,2549	**1,15**	0,3749	**1,61**	0,4463	**2,14**	0,4838	**4,50**	0,499997
0,24	0,0948	**0,70**	0,2580	**1,16**	0,3770	**1,62**	0,4474	**2,16**	0,4846	**5,00**	0,499997
0,25	0,0987	**0,71**	0,2611	**1,17**	0,3790	**1,63**	0,4484	**2,18**	0,4854		
0,26	0,1026	**0,72**	0,2642	**1,18**	0,3810	**1,64**	0,4495	**2,20**	0,4861		
0,27	0,1064	**0,73**	0,2673	**1,19**	0,3830	**1,65**	0,4505	**2,22**	0,4868		
0,28	0,1103	**0,74**	0,2703	**1,20**	0,3849	**1,66**	0,4515	**2,24**	0,4875		
0,29	0,1141	**0,75**	0,2734	**1,21**	0,3869	**1,67**	0,4525	**2,26**	0,4881		
0,30	0,1179	**0,76**	0,2764	**1,22**	0,3883	**1,68**	0,4535	**2,28**	0,4887		
0,31	0,1217	**0,77**	0,2794	**1,23**	0,3907	**1,69**	0,4545	**2,30**	0,4893		
0,32	0,1255	**0,78**	0,2823	**1,24**	0,3925	**1,70**	0,4554	**2,32**	0,4898		
0,33	0,1293	**0,79**	0,2852	**1,25**	0,3944	**1,71**	0,4564	**2,34**	0,4904		
0,34	0,1331	**0,80**	0,2881	**1,26**	0,3962	**1,72**	0,4573	**2,36**	0,4909		
0,35	0,1368	**0,81**	0,2910	**1,27**	0,3980	**1,73**	0,4582	**2,38**	0,4913		
0,36	0,1406	**0,82**	0,2939	**1,28**	0,3997	**1,74**	0,4591	**2,40**	0,4918		
0,37	0,1443	**0,83**	0,2967	**1,29**	0,4015	**1,75**	0,4599	**2,42**	0,4922		
0,38	0,1480	**0,84**	0,2995	**1,30**	0,4032	**1,76**	0,4608	**2,44**	0,4927		
0,39	0,1517	**0,85**	0,3023	**1,31**	0,4049	**1,77**	0,4616	**2,46**	0,4931		
0,40	0,1554	**0,86**	0,3051	**1,32**	0,4066	**1,78**	0,4625	**2,48**	0,4934		
0,41	0,1591	**0,87**	0,3078	**1,33**	0,4082	**1,79**	0,4633	**2,50**	0,4938		
0,42	0,1628	**0,88**	0,3106	**1,34**	0,4099	**1,80**	0,4641	**2,52**	0,4941		
0,43	0,1664	**0,89**	0,3133	**1,35**	0,4115	**1,81**	0,4649	**2,54**	0,4945		
0,44	0,1700	**0,90**	0,3159	**1,36**	0,4131	**1,82**	0,4656	**2,56**	0,4948		
0,45	0,1736	**0,91**	0,3186	**1,37**	0,4147	**1,83**	0,4664	**2,58**	0,4951		

Приложение 3

Таблица значений $t_\gamma = t(\gamma; n)$

n \ γ	0,95	0,99	0,999	n \ γ	0,95	0,99	0,999
5	2,78	4,60	8,61	20	2,093	2,861	3,883
6	2,57	4,03	6,86	25	2,064	2,797	3,745
7	2,45	3,71	5,96	30	2,045	2,756	3,659
8	2,37	3,50	5,41	35	2,032	2,720	3,600
9	2,31	3,36	5,04	40	2,023	2,708	3,558
10	2,26	3,25	4,78	45	2,016	2,692	3,527
11	2,23	3,17	4,59	50	2,009	2,679	3,502
12	2,20	3,11	4,44	60	2,001	2,662	3,464
13	2,18	3,06	4,32	70	1,996	2,649	3,439
14	2,16	3,01	4,22	80	1,991	2,640	3,418
15	2,15	2,98	4,14	90	1,987	2,633	3,403
16	2,13	2,95	4,07	100	1,984	2,627	3,392
17	2,12	2,92	4,02	120	1,980	2,617	3,374
18	2,11	2,90	3,97	∞	1,960	2,576	3,291
19	2,10	2,88	3,92				

Приложение 4

Таблица значений $q = q\,(\gamma, n)$

γ \ n	0,95	0,99	0,999	γ \ n	0,95	0,99	0,999
5	1,37	2,67	5,64	20	0,37	0,58	0,88
6	1,09	2,01	3,88	25	0,32	0,49	0,73
7	0,92	1,62	2,98	30	0,28	0,43	0,63
8	0,80	1,38	2,42	35	0,26	0,38	0,56
9	0,71	1,20	2,06	40	0,24	0,35	0,50
10	0,65	1,08	1,80	45	0,22	0,32	0,46
11	0,59	0,98	1,60	50	0,21	0,30	0,43
12	0,55	0,90	1,45	60	0,188	0,269	0,38
13	0,52	0,83	1,33	70	0,174	0,245	0,34
14	0,48	0,78	1,23	80	0,161	0,226	0,31
15	0,46	0,73	1,15	90	0,151	0,211	0,29
16	0,44	0,70	1,07	100	0,143	0,198	0,27
17	0,42	0,66	1,01	150	0,115	0,160	0,211
18	0,40	0,63	0,96	200	0,099	0,136	0,185
19	0,39	0,60	0,92	250	0,089	0,120	0,162

Приложение 5

Критические точки распределения χ^2

Число степеней свободы *k*	**Уровень значимости α**					
	0,01	**0,025**	**0,05**	**0,95**	**0,975**	**0,99**
1	6,6	5,0	3,8	0,0039	0,00098	0,00016
2	9,2	7,4	6,0	0,103	0,051	0,020
3	11,3	9,4	7,8	0,352	0,216	0,115
4	13,3	11,1	9,5	0,711	0,484	0,297
5	15,1	12,8	11,1	1,15	0,831	0,554
6	16,8	14,4	12,6	1,64	1,24	0,872
7	18,5	16,0	14,1	2,17	1,69	1,24
8	20,1	17,5	15,5	2,73	2,18	1,65
9	21,7	19,0	16,9	3,33	2,70	2,09
10	23,2	20,5	18,3	3,94	3,25	2,56
11	24,7	21,9	19,7	4,57	3,82	3,05
12	26,2	23,3	21,0	5,23	4,40	3,57
13	27,7	24,7	22,4	5,89	5,01	4,11
14	29,1	26,1	23,7	6,57	5,63	4,66
15	30,6	27,5	25,0	7,26	6,26	5,23
16	32,0	28,8	26,3	7,96	6,91	5,81
17	33,4	30,2	27,6	8,67	7,56	6,41
18	34,8	31,5	28,9	9,39	8,23	7,01
19	36,2	32,9	30,1	10,1	8,91	7,63
20	37,6	34,2	31,4	10,9	9,59	8,26
21	38,9	35,5	32,7	11,6	10,3	8,90
22	40,3	36,8	33,9	12,3	11,0	9,54
23	41,6	38,1	35,2	13,1	11,7	10,2
24	43,0	39,4	36,4	13,8	12,4	10,9
25	44,3	40,6	37,7	14,6	13,1	11,5
26	45,6	41,9	38,9	15,4	13,8	12,2
27	47,0	43,2	40,1	16,2	14,6	12,9
28	48,3	44,5	41,3	16,9	15,3	13,6
29	49,6	45,7	42,6	17,7	16,0	14,3
30	50,9	47,0	43,8	18,5	16,8	15,0

Приложение 6

Критические точки распределения Стьюдента

Число степеней свободы k	Уровень значимости α (двусторонняя критическая область)					
	0,10	**0,05**	**0,02**	**0,01**	**0,002**	**0,001**
1	6,31	12,7	31,82	63,7	318,3	637,0
2	2,92	4,30	6,97	9,92	22,33	31,6
3	2,35	3,18	4,54	5,84	10,22	12,9
4	2,13	2,78	3,75	4,60	7,17	8,61
5	2,01	2,57	3,37	4,03	5,89	6,86
6	1,94	2,45	3,14	3,71	5,21	5,96
7	1,89	2,36	3,00	3,50	4,79	5,40
8	1,86	2,31	2,90	3,36	4,50	5,04
9	1,83	2,26	2,82	3,25	4,30	4,78
10	1,81	2,23	2,76	3,17	4,14	4,59
11	1,80	2,20	2,72	3,11	4,03	4,44
12	1,78	2,18	2,68	3,05	3,93	4,32
13	1,77	2,16	2,65	3,01	3,85	4,22
14	1,76	2,14	2,62	2,98	3,79	4,14
15	1,75	2,13	2,60	2,95	3,73	4,07
16	1,75	2,12	2,58	2,92	3,69	4,01
17	1,74	2,11	2,57	2,90	3,65	3,95
18	1,73	2,10	2,55	2,88	3,61	3,92
19	1,73	2,09	2,54	2,86	3,58	3,88
20	1,73.	2,09	2,53	2,85	3,55	3.85
21	1.72	2,08	2,52	2,83	3.53	3.82
22	1,72	2,07	2,51	2.82	3,51	3.79
23	1.71	2.07	2.50	2,81	3,49	3.77
24	1.71	2,06	2,49	2.80	3,47	3.74
25	1,71	2,06	2,49	2,79	3,45	3.72
26	1,71	2,06	2,48	2,78	3,44	3.7!
27	1,71	2,05	2,47	2,77	3,42	3.69
28	1,70	2,05	2,46	2,76	3,40	3.66
29	1,70	2,05	2,46	2,76	3,40	3,66
30	1,70	2,04	2,46	2,75	3,39	3.65
40	1,68	2,02	2,42	2.70	3,31	3.55
60	1,67	2,00	2,39	2,66	3,23	3,46
120	1,66	1,98	2,36	2,62	3,17	3.37
∞	1,64	1,96	2,33	2,58	3,09	3,29
	0,05	**0,025**	**0,01**	**0,005**	**0,001**	**0,0005**
	Уровень значимости α (односторонняя критическая область)					

Приложение 7

Критические точки распределения *F* Фишера – Снедекора

k_1 – число степеней свободы большей дисперсии

k_2 – число степеней свободы меньшей дисперсии

Уровень значимости α = 0,01												
k_2 \ k_1	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	4052	4999	5403	5625	5764	5889	5928	5981	6022	6056	6082	6106
2	98,49	99,01	90,17	99,25	99,33	99,30	99,34	99,36	99,36	99,40	99,41	99,42
3	34,12	30,81	29,46	28,71	28,24	27,91	27,67	27,49	27,34	27,23	27,13	27,05
4	21,20	18,00	16,69	15,98	15,52	15,21	14,98	14,80	14,66	14,54	14,45	14,37
5	16,26	13,27	12,06	11,39	10,97	10,67	10,45	10,27	10,15	10,05	9,96	9,89
6	13,74	10,92	9,78	9,15	8,75	8,47	8,26	8,10	7,98	7,87	7,79	7,72
7	12,25	9,55	8,45	7,85	7,46	7,19	7,00	6,84	6,71	6,62	6,54	6,47
8	11,26	8,65	7,59	7,01	6,63	6,37	6,19	6,03	5,91	5,82	5,74	5,67
9	10,56	8,02	6,99	6,42	6,06	5,80	5,62	5,47	5,35	5,26	5,18	5,11
10	10,04	7,56	6,55	5,99	5,64	5,39	5,21	5,06	4,95	4,85	4,78	4,71
11	9,86	7,20	6,22	5,67	5,32	5,07	4,88	4,74	4,63	4,54	4,46	4,40
12	9,33	6,93	5,95	5,41	5,06	4,82	4,65	4,50	4,39	4,30	4,22	4,16
13	9,07	6,70	5,74	5,20	4,86	4,62	4,44	4,30	4,19	4,10	4,02	3,96
14	8,86	6,51	5,56	5,03	4,69	4,46	4,28	4,14	4,03	3,94	3,86	3,80
15	8,68	6,36	5,42	4,89	4,56	4,32	4,14	4,00	3,89	3,80	3,73	3,67
16	8,53	6,23	5,29	4,77	4,44	4,20	4,03	3,89	3,78	3,69	3,61	3,55
17	8,40	6,11	5,18	4,67	4,34	4,10	3,93	3,79	3,68	3,59	3,52	3,45
Уровень значимости α = 0,05												
k_2 \ k_1	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**
1	161	200	216	225	230	234	237	239	241	242	243	244
2	18,51	19,00	19,16	19,25	19,30	19,33	19,36	19,37	19,38	19,39	19,40	19,41
3	10,13	9,55	9,28	9,12	9,01	8,94	8,88	8,84	8,81	8,78	8,76	8,74
4	7,71	6,94	6,59	6,39	6,26	6,16	6,09	6,04	6,00	5,96	5,93	5,91
5	6,61	5,79	5,41	5,19	5,05	4,95	4,88	4,82	4,78	4,74	4,70	4,68
6	5,99	5,14	4,76	4,53	4,39	4,28	4,21	4,15	4,10	4,06	4,03	4,00
7	5,59	4,74	4,35	4,12	3,97	3,87	3,79	3,73	3,68	3,63	3,60	3,57
8	5,32	4,46	4,07	3,84	3,69	3,58	3,50	3,44	3,39	3,34	3,31	3,28
9	5,12	4,26	3,86	3,63	3,48	3,37	3,29	3,23	3,18	3,13	3,10	3,07
10	4,96	4,10	3,71	3,48	3,33	3,22	3,14	3,07	3,02	2,97	2,94	2,91
11	4,84	3,98	3,59	3,36	3,20	3,09	3,01	2,95	2,90	2,86	2,82	2,79
12	4,75	3,88	3,49	3,26	3,11	3,00	2,92	2,85	2,80	2,76	2,72	2,69
13	4,67	3,80	3,41	3,18	3,02	2,92	2,84	2,77	2,72	2,67	2,63	2,60
14	4,60	3,74	3,34	3,11	2,96	2,85	2,77	2,70	2,65	2,60	2,56	2,53
15	4,54	3,68	3,29	3,06	2,90	2,79	2,70	2,64	2,59	2,55	2,51	2,48
16	4,49	3,63	3,24	3,01	2,85	2,74	2,66	2,59	2,54	2,49	2,45	2,42
17	4,45	3,59	3,20	2,96	2,81	2,70	2,62	2,55	2,50	2,45	2,41	2,38