1. ОСНОВНЫЕ ПОНЯТИЯ СТАТИСТИЧЕСКОЙ ОБРАБОТКИ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

1.1. Цели, задачи и методы статистической обработки

Экспериментальные данные, получаемые в процессе измерений, являются основным источником информации о характеристиках процессов, протекающих в изучаемых объектах. Причем, чем менее строгой является наука, тем более значимую роль в ней играет эксперимент. В науках, использующих развитый математический аппарат (например, в физике), многие результаты могут быть получены и обоснованы теоретически на базе предшествующих знаний, тогда как в биологии или медицине эксперимент зачастую является единственным способом подтверждения справедливости гипотезы и результатов теоретического исследования.

Научный эксперимент решает следующие основные задачи:

- выступает средством получения новых научных данных;
- является способом выделения общего в серии сходных явлений, обоснования закономерностей, формирования гипотез;
- выступает средством проверки гипотез и теорий, критерием их истинности.

Этапами постановки и проведения эксперимента являются:

- определение цели и задач исследования,
- выбор конкретных методик,
- непосредственное проведение эксперимента,
- обработка данных эксперимента.

Рассмотрим следующую модель медико-биологического эксперимента.

Целью эксперимента является обоснование изменения состояния экспериментальной группы животных в результате целенаправленного воздействия на нее нового лекарственного средства.

Для того, чтобы выделить в явном виде результат целенаправленного воздействия, необходимо взять аналогичную контрольную группу и посмотреть, что происходит с ней в отсутствии воздействий или при использовании традиционного лекарственного средства.

Поэтому методика эксперимента, схематически изображенная на рис. 1, заключается в следующем:

- на основании сравнения I установить совпадение начальных состояний экспериментальной и контрольной группы;

- реализовать воздействие на экспериментальную (III) и контрольную (IV) группы;

- на основании сравнения II установить различие конечных состояний экспериментальной и контрольной групп.

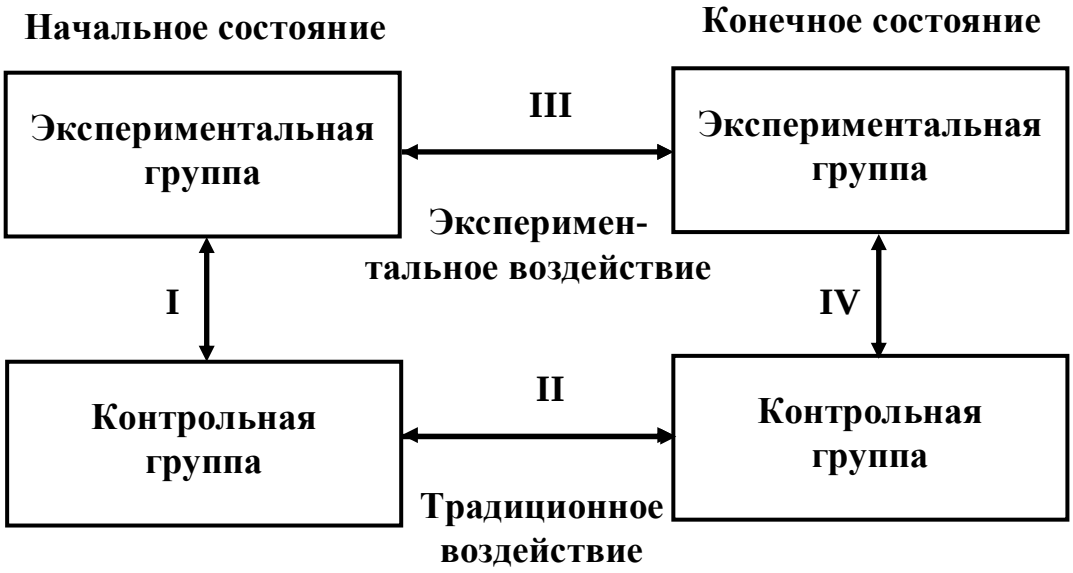


Рис. 1. Структура эксперимента

Следующим этапом является обработка экспериментальных данных, основной целью которой является нахождение максимального количества содержащейся в них и необходимой экспериментатору информации с максимальной достоверностью и точностью.

При проведении эксперимента в реальных условиях и с помощью реальных измерительных систем получаемые результаты всегда носят вероятностный характер вследствие неизбежного влияния внешних факторов, неполноты знаний о состоянии объекта исследования и т.д.

Таблица 1.

Задачи статистического анализа и методы их решения

Задачи статистического анализа данных	Методы статистического анализа данных
Описание данных	Описательная статистика
Изучение сходства или различия данных	Параметрические и непараметрические статистические

	критерии
Исследование зависимостей	Корреляционный анализ, регрессионный анализ, дисперсионный анализ
Снижение размерности	Метод главных компонент, факторный анализ, многомерное шкалирование
Классификация и прогноз	Дискриминантный анализ, кластерный анализ

Поэтому при подведении итогов эксперимента существенную роль играют методы теории вероятности и математической статистики, которые дают возможность решать следующие задачи:

- компактно и информативно описывать результаты эксперимента;
- устанавливать степень достоверности сходства и различия исследуемых объектов на основании результатов измерений их показателей;
- анализировать наличие или отсутствие зависимости между различными показателями (явлениями);
- количественно описывать эти зависимости;
- классифицировать изучаемые объекты и прогнозировать значения их показателей и характеристик, и др.

В табл. 1 представлены основные статистические методы решения вышеуказанных задач.

1.2 Шкалы измерений

Количество информации и применяемые при обработке методы анализа зависят, в первую очередь, от типа шкалы, в которой данные получены.

Шкала – это числовая система, в которой отношения между различными свойствами изучаемых явлений, процессов переведены в свойства множества чисел.

Основным классификационным признаком является **мощность шкалы**, т.е. возможность допустимых преобразований шкалы, которые не меняют соотношений между объектами измерения.

Например, при измерении длины переход от дюймов к метрам не меняет соотношений между длинами рассматриваемых объектов. Если

первый объект длиннее второго, то это будет установлено и при измерении в дюймах, и при измерении в метрах. При этом численное значение длины в дюймах отличается от численного значения длины в метрах, не меняется лишь результат сравнения длин двух объектов.

В зависимости от мощности различают, как показано на рисунке 2, шкалы отношений, интервальные шкалы, порядковые (ранговые) шкалы и номинальные шкалы (шкалы наименований).

Рисунок 2. Классификация шкал измерений

Чем большую мощность имеет шкала, тем больше информативных характеристик о результатах эксперимента можно получить, т.е. тем выше достоверность проведенных исследований. В идеальном случае результаты эксперимента должны быть представлены в шкале отношений.

К шкалам **качественных признаков** относятся порядковая шкала и шкала наименований.

Шкала наименований (номинальная шкала) фактически уже не связана с понятием «величина» и используется только с целью отличить один объект от другого: номер животного в группе или присвоенный ему уникальный шифр и т.п.

В шкале наименований, например, измеряются значения признака «пол»: «мужской» и «женский». Эти классы будут различимы независимо от того, какие различные термины или знаки для их обозначений будут использованы: “female” и “male”, или “А” и “Б” и т.д. Следовательно, для шкалы наименований применимы любые взаимно-однозначные преобразования, то есть сохраняющие четкую различимость объектов. Таким образом, самая слабая шкала, шкала наименований, допускает самый широкий диапазон преобразований.

Порядковая шкала (шкала рангов) – шкала, относительно значений которой уже нельзя говорить ни о том, во сколько раз измеряемая величина больше (меньше) другой, ни на сколько она

больше (меньше). Такая шкала только упорядочивает объекты, приписывая им те или иные баллы.

Например, так построена шкала твердости минералов Мооса: взят набор 10 эталонных минералов для определения относительной твердости методом царапанья. За 1 принят тальк, за 2 – гипс, за 3 – кальцит и так далее до 10 – алмаз. Любому минералу соответственно однозначно может быть приписана определенная твердость. Если исследуемый минерал, допустим, царапает кварц (7), но не царапает топаз (8), то соответственно его твердость будет равна 7. Аналогично построены шкалы силы ветра Бофорта и землетрясений Рихтера.

Шкалы порядка широко используются в педагогике, психологии, медицине и других науках, не столь точных, как физика или химия. В частности, повсеместно распространенная шкала школьных отметок в баллах может быть отнесена к шкале порядка. В медико-биологических исследованиях шкалы порядка встречаются, например, при анализе свертывания крови (тромботест). Отличие порядковой шкалы (шкалы рангов) от шкалы наименований заключается в том, что в шкале рангов классы (группы) объектов упорядочены. Поэтому произвольным образом изменять значения признаков нельзя – должна сохраняться упорядоченность объектов (порядок следования одних объектов за другими). Следовательно, для порядковой шкалы допустимым является любое монотонное преобразование. Например, если проба тромботеста у животного А 5 баллов, а у животного Б 4 балла, то их упорядочение не изменится, если мы число баллов умножим на одинаковое для всех животных положительное число, или сложим с некоторым одинаковым для всех числом, или возведем в квадрат и т.д.

Частным случаем порядковой шкалы является **дихотомическая шкала**, в которой имеются всего две упорядоченные градации, например, «выжил после эксперимента», «не выжил».

Шкала интервалов характеризуется тем, что для нее не существует ни естественного начала отсчета, ни естественной единицы измерения.

Примером шкалы интервалов является шкала температур по Цельсию, Реомюру или Фаренгейту. Шкала Цельсия, как известно, была установлена следующим образом: за ноль была принята точка замерзания воды, за 100 градусов – точка ее кипения, и, соответственно, интервал температур между замерзанием и кипением воды поделен на 100 равных частей.

Для шкалы интервалов допустимо уже не любое монотонное

преобразование, а только такое, которое сохраняет отношение разностей оценок, то есть линейное преобразование – умножение на положительное число и добавление постоянного числа. Например, если к значению температуры в градусах Цельсия добавить 273˚С, то получим температуру по Кельвину, причем разности любых двух температур в обеих шкалах буду одинаковы. А если от шкалы Цельсия перейти к шкале Фаренгейта, то для любых четырех температур отношение разности первой и второй к разности третьей и четвертой будут одинаковы в обеих шкалах.

Шкала отношений – самая мощная шкала и наиболее распространенная шкала. Она позволяет оценивать, во сколько раз один измеряемый объект больше (меньше) другого объекта, принимаемого за эталон, единицу. Для шкал отношений существует естественное начало отсчета (нуль), но нет естественной единицы измерений.

Шкалами отношений измеряются почти все физические величины – время, линейные размеры, площади, объемы, сила тока, мощность и т.д. В медико-биологических исследованиях шкала отношений используется, например, когда измеряется время появления того или иного признака после начала воздействия, интенсивность воздействия до появления какого-либо признака, концентрации веществ, временные показатели электрокардиограммы и т.п.).

В шкале отношений возможны лишь преобразования подобия - умножения на положительное число. Это означает, что, например, отношение масс двух предметов не зависит от того, в каких единицах измерены массы – граммах, килограммах и т.д.

Шкалы интервалов и отношений являются шкалами **количественных признаков**.

Таблица 2. Шкалы и допустимые преобразования

Шкала	Допустимое преобразование
Наименований	Взаимно-однозначное
Порядковая	Строго возрастающее
Интервальная	Линейное
Отношений	Подобия

В таблице 2 приведено соответствие между шкалами и допустимыми преобразованиями.

В процессе развития соответствующей области знания тип шкалы

может меняться. Так, сначала температура измерялась по интервальной шкале, а после открытия абсолютного нуля температуру можно считать измеренной по шкале отношений (шкала Кельвина).

1.3. Понятие генеральной совокупности и выборки

Совокупность объектов или наблюдений, все элементы которой подлежат изучению при статистическом анализе, называется **генеральной совокупностью.**

Исследование всего набора элементов генеральной совокупности часто оказывается невозможным или нецелесообразным. Например, изучение эффективности нового лекарственного средства на всех больных, для которых это средство разработано, или определение предела прочности всего объема выпуска деталей заводом, для чего необходимо было бы разрушить эти детали.

В таких случаях рассматривают некоторую часть генеральной совокупности, которая называется **выборочной совокупностью** или **выборкой**.

Сущность выборочного метода в математической статистике заключается в том, чтобы по определенной части генеральной совокупности (выборке) судить о ее свойствах в целом.

Для этого выборка должна быть представительной (репрезентативной).

Репрезентативность - это способность выборочной совокупности как количественно, так и качественно отражать свойства генеральной совокупности.

Репрезентативность выборки обеспечивается выполнением четырех требований: случайный отбор, однородность, независимость, достаточный объем.

1. Основное требование, предъявляемое к формированию выборки - **рандомизированный отбор**, т.е. случайный отбор элементов выборки из генеральной совокупности, при котором каждой единице наблюдения обеспечивается равная вероятность попадания в выборку.

Случайный отбор может производиться **непосредственно из генеральной совокупности**. При этом случайность достигается путем применения жеребьевки или использования таблицы случайных чисел. Другим способом случайного отбора является **механический отбор**, когда генеральная совокупность разбивается на равные части, из которых затем в заранее обусловленном порядке отбирают единицы

наблюдения под определенным номером, так, чтобы обеспечить необходимое число наблюдений.

При случайном отборе возможны два варианта формирования выборки:

- **повторная выборка**, когда каждый элемент, случайно отобранный и исследованный, возвращается в общую совокупность и может быть отобран повторно;

- **бесповторная выборка**, когда отобранный элемент не возвращается в общую совокупность.

Кроме вышеуказанных способов при формировании выборки используют:

- т**ипический отбор**, при осуществлении которого генеральная совокупность делится по некоторому признаку на типические группы, и отбор единиц производится из типических групп;

- с**ерийный отбор** с равновеликими сериями состоит в выборе не единиц совокупности, а некоторых групп совокупностей одинаковых объемов (серий);

- **комбинированный отбор**, при осуществлении которого комбинируются различные методы. Сначала, например, отбираются серии, а затем из отобранных серий производится индивидуальная выборка единиц.

2. **Однородность** выборки предполагает, что выборка должна состоять из элементов, принадлежащих одному объекту и выполненных одинаковым способом измерения. Это требование часто трудновыполнимо, поэтому существуют специальные статистические методы проверки однородности.

3. **Независимость** предполагает, что результаты каждого наблюдения или измерения не зависят от результатов последующих или предыдущих наблюдений.

4. **Достаточный объем** выборки определяется, исходя из ошибки репрезентативности, которая показывает, на сколько отличаются характеристики выборочной совокупности от соответствующих характеристик генеральной совокупности.

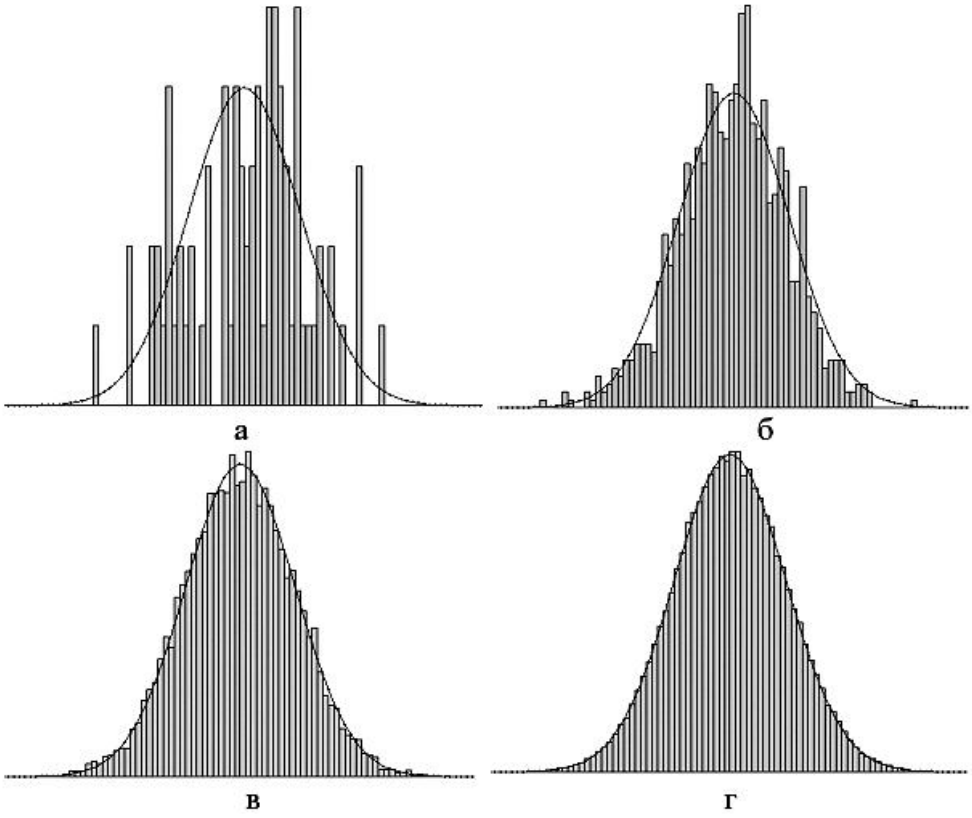


Рис. 3. Гистограммы эмпирических распределений для различного количества элементов выборки из нормально распределенной генеральной совокупности:
а) n = 100, б) n = 1000, в) n = 10000, г) n = 100000

Взаимосвязь статистических показателей выборочной и генеральной совокупностей определяется законом больших чисел, выражаясь в теореме П. Л.Чебышева: чем больше число некоторых случайных величин, тем их средняя арифметическая ближе к средней арифметической генеральной совокупности, т.е. тем меньше разница между показателями выборочной и генеральной совокупностей. По мере увеличения числа наблюдений вероятность осуществления приближения показателя выборки к показателю генеральной совокупности становится все больше, стремясь к единице, если число наблюдений стремится к бесконечности.

На рисунке 3 показано, как изменяется вид гистограмм, построенных по выборочным данным из одной генеральной совокупности при увеличении количества элементов выборки.

Эмпирическим путем установлено, что надежность статистических оценок резко снижается в диапазоне от 60 до 10 – 20 наблюдений, а при меньшем числе наблюдений применять статистический анализ в большинстве случаев нецелесообразно. При

малых выборках прибегают к специальным способам расчетов.

1.4 Законы распределения вероятностей, используемые при анализе данных

При оценивании экспериментальных данных и проверке статистических гипотез находит применение ряд теоретических законов распределения.

Наиболее важным из них является нормальное распределение. С ним связаны распределения хи-квадрат, Стьюдента, Фишера. Для указанных законов значения функций распределения находятся по статистическим таблицам (Приложение 1, 2) или с использованием стандартных процедур пакетов прикладных статистических программ.

Нормальное распределение.

Этот вид распределения является наиболее распространенным в связи с центральной предельной теоремой теории вероятностей, согласно которой распределение суммы независимых случайных величин стремится к нормальному с увеличением их количества при произвольном законе распределения отдельных слагаемых, если слагаемые обладают конечной дисперсией.

Так как реальные физические явления часто представляют собой результат суммарного воздействия многих факторов, то в таких случаях нормальное распределение является хорошим приближением наблюдаемых значений. Распределение многих статистик является нормальным или может быть получено из нормальных с помощью некоторых преобразований.

Можно сказать, что нормальное распределение представляет собой одну из эмпирически проверенных истин относительно общей природы действительности и может рассматриваться как один из фундаментальных законов природы.

Случайная величина ξ имеет нормальное распределение вероятностей, если ее плотность определяется по приведенной формуле:

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

где m – математическое ожидание, σ - стандартное отклонение.

На рисунке 4 приведены кривые плотности вероятности нормального распределения с m = 20 и разными значениями σ.

Точная форма нормального распределения (характерная

"колоколообразная кривая") определяется двумя параметрами: средним и стандартным отклонением. Характерное свойство нормального распределения состоит в том, что 68% всех его наблюдений лежат в диапазоне ±1 стандартное отклонение от среднего, а диапазон ±2 стандартных отклонения содержит 95% значений. При нормальном распределении практически все наблюдения (т.е. более 99.99%) попадут в диапазон ±3 стандартных отклонения.

Стандартное нормальное распределение, которое имеет среднее 0 и стандартное отклонение 1, используется при проверке различных гипотез, в том числе о среднем значении, о различии между двумя средними.

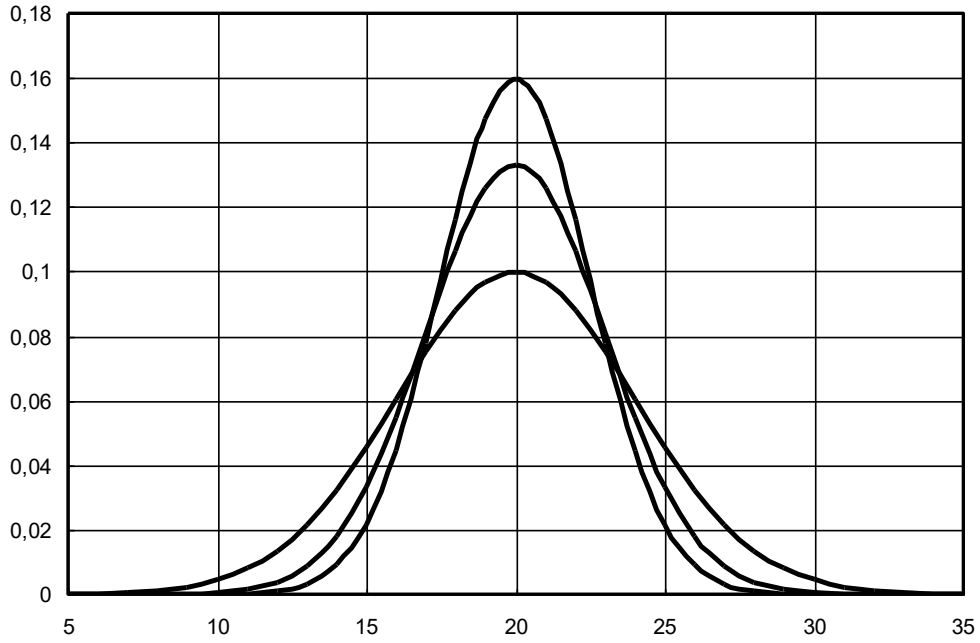


Рисунок 4. Плотность вероятности нормального распределения при разных значениях дисперсии.

При операциях с нормально распределенными случайными величинами в процессе анализа данных возникает несколько новых видов распределений, **связанных с нормальным.**

Распределение Стьюдента.

Пусть случайные величины $\xi_0,\xi_1,\ldots\xi_n$ независимы и каждая из них имеет стандартное нормальное распределение N(0,1). Введем следующую случайную величину:

$$t_n = \frac{\xi_0}{\sqrt{\frac{1}{\nu}\sum_{i=1}^{n}\xi_i^2}}$$

где ν - число степеней свободы.

Распределение этой величины называется распределением Стьюдента или t-распределением.

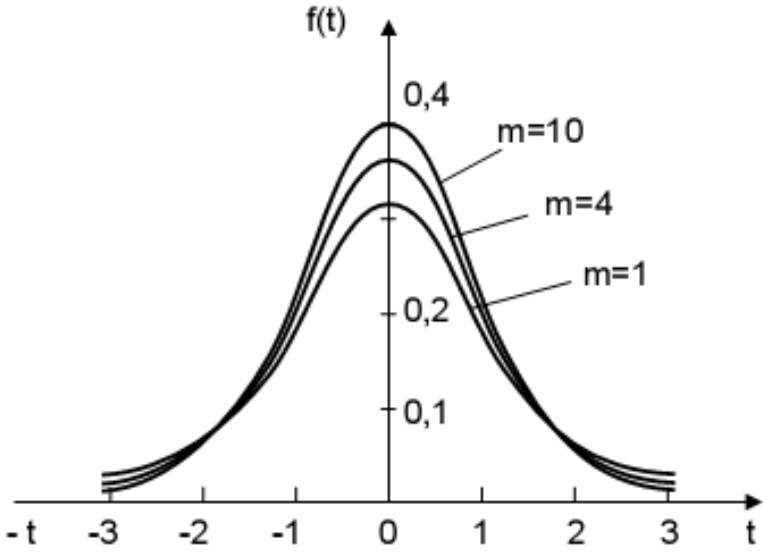


Рисунок 5. Плотность распределения Стьюдента для различных значений степеней свободы ν.

На рисунке 5 показана плотность распределения Стьюдента по сравнению с нормальным распределением.

В математической статистике **степень свободы** - наименьшее число независимых (свободных) величин в данной задаче.

Если n - число величин, m - число ограничений (связей), то число степеней свободы ν = n - m.

Так, если сумма трех данных равна 8, то первые два из них могут принимать любые значения, но если они определены, то третье значение становится автоматически известным. Если, например, значение первого данного равно 3, а второго -1, то третье может быть равным только 4. Таким образом, в такой выборке имеются только две степени свободы. В общем случае для выборки в n данных существует n-1 степень свободы

Например, знаменатель в формуле выборочной дисперсии всегда равен разности между объёмом выборки и числом связей, наложенных на эту выборку. Эта разность фактически показывает, какое количество элементов выборки можно произвольно изменять, не нарушая связей.

Если имеются две независимые выборки, то число степеней свободы для первой из них составляет ν_1-1, а для второй — ν_2-1. А

поскольку при определении достоверности разницы между ними опираются на анализ каждой выборки, число степеней свободы, по которому нужно будет находить критерий t в статистической таблице, будет составлять ($\nu_1 + \nu_2$) - 2.

Если же речь идет о двух зависимых выборках, то в основе расчета лежит вычисление суммы разностей, полученных для каждой пары результатов (т.е., например, разностей между результатами до и после воздействия на одного и того же испытуемого). Поскольку одну (любую) из этих разностей можно вычислить, зная остальные разности и их сумму, число степеней свободы для определения критерия t будет равно n-1.

Форма распределения Стьюдента зависит от числа степеней свободы.

Математическое ожидание его равно 0, дисперсия – n/(n-2).

При увеличении этого параметра кривая плотности распределения сужается и при $n \to \infty$ асимптотически приближается к нормальному распределению.

Распределение Стьюдента применяется для описания ошибок выборки при $n \geq 30$. При n>100 данное распределение практически соответствует нормальному, для $30 < n < 100$ различия между распределением Стьюдента и нормальным распределением составляют несколько процентов. Поэтому относительно оценки ошибок малыми считаются выборки объемом не более 30 единиц, большими – объемом более 100 единиц.

Распределение хи-квадрат

Пусть случайные величины $\xi_0, \xi_1, \ldots \xi_\nu$ независимы и каждая из них имеет стандартное нормальное распределение N(0,1). Введем следующую случайную величину:

$$\chi_n^2 = \xi_1^2 + \ldots + \xi_\nu^2,$$

где ν - число степеней свободы.

Полученное распределение называется распределением хи-квадрат. Форма его также зависит от числа степеней свободы.

Математическое ожидание его равно n, дисперсия – 2n.

На рисунке 6 приведено измерение плотности вероятностей распределения хи-квадрат при степенях свободы 2,4,10.

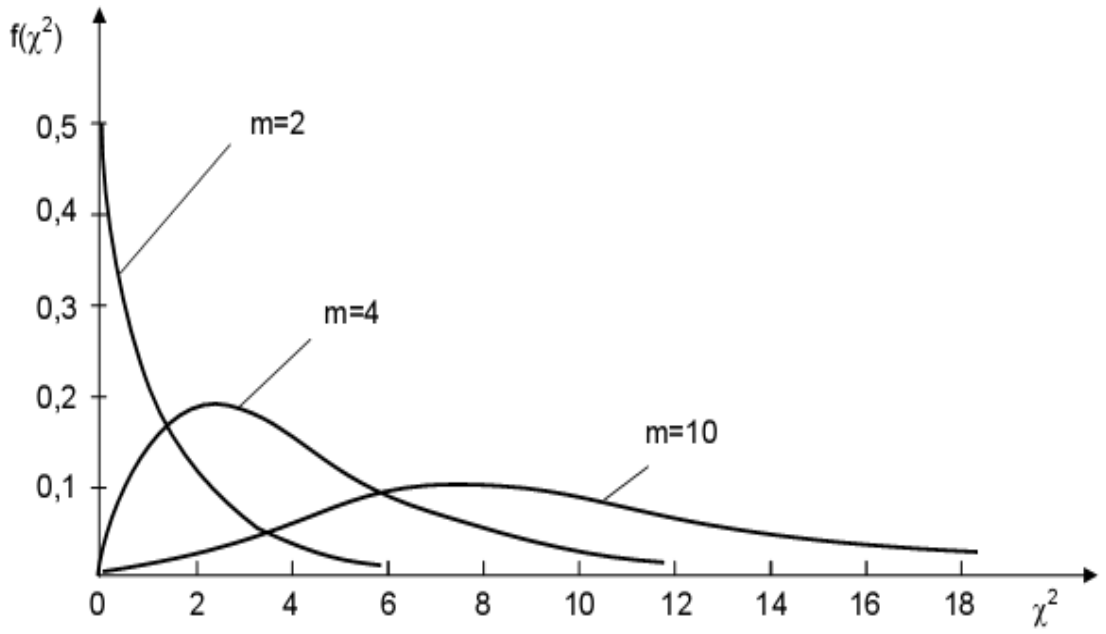


Рис. 6. Плотность распределения хи-квадрат для различных значений степеней свободы ν.

Распределение Фишера (F-распределение **.**

Пусть случайные величины $\xi_0,\xi_1,\ldots)_n$ и $\eta_0,\eta_1,\ldots\eta_m$ независимы и каждая из них имеет стандартное нормальное распределение N(0,1).

Тогда случайная величина

$$F_{m,n} = \frac{n(\eta_1^2 + \ldots + \eta_m^2)}{m(\xi_1^2 + \ldots\xi_n^2)}$$

имеет распределение Фишера с параметрами m и n. Плотность распределения вероятностей зависит от соотношения этих параметров.

F-распределение является асимметричным и обычно используется в дисперсионном анализе.

2. СТАТИСТИЧЕСКОЕ ОПИСАНИЕ ЧИСЛОВЫХ ХАРАКТЕРИСТИК ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Все характеристики, получаемые для выборки из генеральной совокупности, называют в математической статистике выборочными (например, выборочное среднее, выборочное или эмпирическое распределение и т. д.).

Пусть в результате эксперимента получена выборка данных x = (x_1, x_2 . . . , x_n), т.е. набор из n независимых и одинаково распределенных

случайных величин. Если элементы выборки упорядочить по возрастанию, получится новая статистическая совокупность: $x_1 \le x_2 \le \ldots \le x_n$.

Статистическая совокупность, расположенная в порядке возрастания элементов выборки, называется **вариационным рядом** или **ранжированной** выборкой, а ее различные (несовпадающие) элементы - **вариантами**.

Статистическим распределением выборки называется соотношение между значениями вариант и соответствующими им частотами n_i или относительными частотами $P_i = n_i / n$. Табличный вариант записи параметров статистического распределения показан в таблице 3.

Таблица 3.

Варианты x_i	x_1	x_2	x_3	...	x_i	...	x_k
Число наблюдений (частота) n_i	n_1	n_2	n_3	...	n_i	...	n_k
Относительная частота P_i	P_1	P_2	P_3	...	P_i	...	P_k

Эмпирической функцией распределения (функцией распределения выборки) называют функцию $F_э(x)$, определяющую для каждого значения x относительную частоту события ($X < x$):

$$F_э(x) = \frac{n_x}{n},$$

где n_x - число вариант, меньших x; n - объем выборки.

Эмпирическим, или выборочным аналогом плотности распределения является гистограмма.

Гистограммой называется столбчатая диаграмма, по оси абсцисс которой откладываются сгруппированные элементы выборки, а по оси ординат - соответствующая им частота n_i или относительная частота P_i.

Величина интервала группировки существенно влияет на общий вид гистограммы. Если длина интервала мала, то преобладает влияние случайных колебаний. При слишком больших интервалах скрадываются характерные черты распределения.

Число интервалов группировки принято определять по формуле Стерджесса:

$k = k(n) = 1 + [3.322 \lg n]$,

В соответствии с формулой Стерджесса при n = 50 - 100

минимальное число интервалов группировки обычно принимается 6 - 8, а при n > 100 - не менее 10 - 15.

После того, как определены границы всех интервалов, остается распределить по этим интервалам выборочные варианты. Если какая-либо из вариант попадает точно на границу соседних интервалов группировки, то такие варианты могут быть отнесены к любому из соседних интервалов по выбору.

Вариационные ряды и гистограммы дают наглядное представление о том, как варьируют экспериментальные данные в выборочной совокупности. Но они недостаточны для полной характеристики выборки, поскольку содержат много деталей, охватить которые невозможно без применения обобщающих числовых характеристик.

Числовые характеристики выборки дают количественное представление об эмпирических данных и позволяют сравнивать их между собой.

Для результатов измерений в шкале отношений числовые показатели описательной статистики можно разбить на несколько групп: характеристики положения, рассеяния и формы эмпирических распределений.

Характеристики положения описывают положение экспериментальных данных на числовой оси. Примеры таких данных - максимальный и минимальный элементы выборки, среднее значение, медиана, мода и др.

Среднее или среднее арифметическое - такое значение признака, сумма отклонений от которого выборочных значений признака равна нулю (с учетом знака отклонения). Среднее является моментом первого порядка.

Если воспользоваться геометрической интерпретацией, то среднее арифметическое можно определить как точку на оси x, которая является абсциссой центра масс гистограммы.

Среднее, как и другие числовые характеристики выборки, может вычисляться как по необработанным первичным данным, так и по результатам группировки этих данных.

Для несгруппированных данных среднее определяется по следующей формуле:

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i;$$

где n — объем выборки; x_i — варианты выборки.

Если данные сгруппированы, то

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{k} n_i x_i;$$

где n - объем выборки; k - число интервалов группировки; n_i - частота i-ого интервала; x_i - срединное значение i-ого интервала.

Среднее, вычисленное по сгруппированным данным, называется **взвешенным средним**, т.к. в данном случае x_i суммируются с коэффициентами (**весами**), равными частотам попадания в интервалы группировки.

Медианой называется значение исследуемого признака, справа и слева от которого находится одинаковое число упорядоченных элементов выборки.

Для вычисления медианы выборку ранжируют, т. е. располагают данные в порядке возрастания или убывания, и в ранжированной выборке, содержащей n членов, ранг R (порядковый номер) медианы определяется как:

$$R = \frac{n+1}{2}.$$

Отличительные особенности медианы:

- ее значение не зависит от формы распределения эмпирических данных;

- если эмпирическое распределение имеет симметричную форму, медиана совпадает со средним значением;

- когда эмпирическое распределение оказывается сильно асимметричным, медиана представляет собой лучшую характеристику центра распределения, а среднее теряет свою практическую ценность, поскольку при этом большая часть данных оказывается выше или ниже среднего.

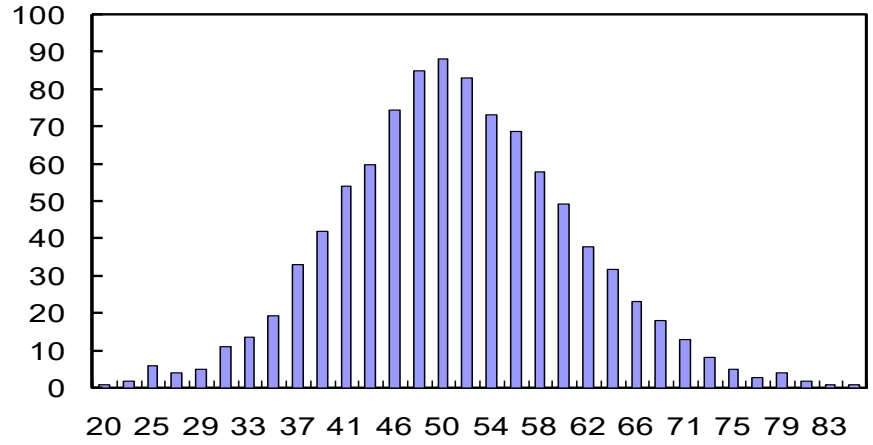
На рисунке 5а) показана гистограмма симметричного распределения, среднее и медиана совпадают и равны 50.

На рисунке 5б) с асимметричным распределением и средним 50 медиана равна 40 и, как видно, более точно отражает центр группирования выборки.

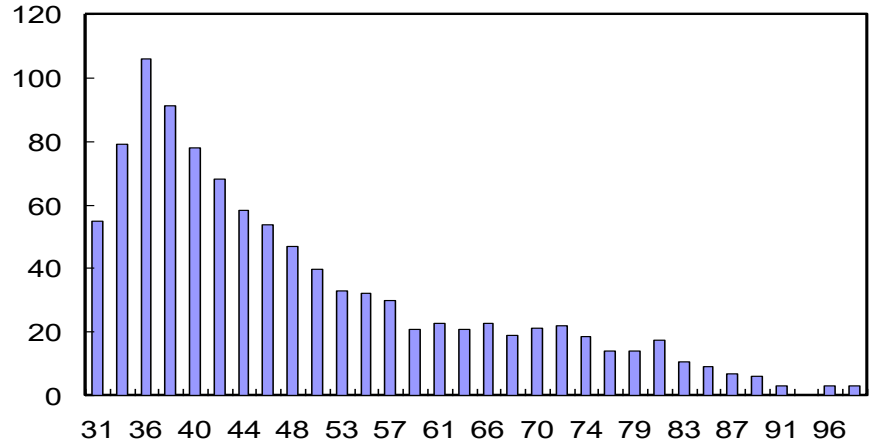
Для более детального описания формы эмпирического распределения используют выборочные квантили.

Выборочной квантилью порядка α называется функция выборки, равная элементу вариационного ряда с номером $[n\alpha+1]$).

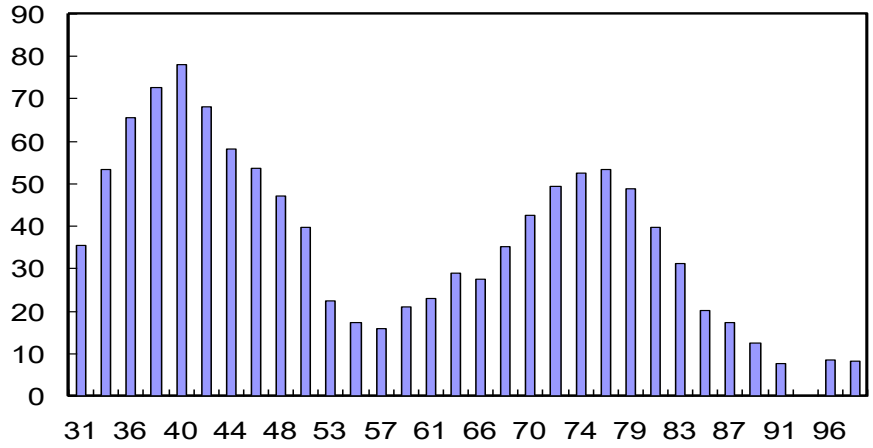
При α = 0.5 квантиль является **медианой**, квантили с α = 0.1, 0.2, ..., 0.9 называются **децилями**, квантили с α = 0.25, 0.5, 0.75 называются **квартилями**.



а)



б)



в)

Рисунок 5. Примеры гистограмм: а) симметричного и б) асимметричного унимодальных распределений, (в) полимодального распределения.

Модой называется такое значение измеренного признака, которым обладает максимальное число элементов выборки, то есть значение, которое встречается в выборке наиболее часто.

Ряд называется **унимодальным**, если в нем только одно модальное значение и **полимодальным**, если есть несколько значений признака, которые встречаются одинаково часто. Пример полимодального распределения показан на рисунке 5в).

Полимодальное распределение может быть следствием того, что выборка экспериментальных данных получена из разных генеральных совокупностей или нарушены условия репрезентативности. Например, двумодальное распределение может быть получено, если при анализе количества гемоглобина в крови выборка будет составлена без учета пола пациентов, тогда как известно, что средний уровень гемоглобина у женщин ниже, чем у мужчин.

Для одномодальных распределений, симметричных относительно среднего значения, мода совпадает со средним значением и медианой, для асимметричных среднее значение обычно смещено относительно моды в сторону более длинного "хвоста" изменения плотности вероятностей.

Характеристики рассеяния описывают степень разброса данных относительно своего центра (среднего значения). К ним относятся: дисперсия, стандартное отклонение, коэффициент вариации и др.

Размах или разность между максимальной и минимальной вариантами выборки определяется как:

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

Информативность этого показателя невелика. Распределения могут сильно отличаться по форме по форме, но иметь одинаковый размах. Размах используется иногда в практических исследованиях при малых объемах выборки для приблизительной оценки вариации данных.

Дисперсия или момент второго порядка является важнейшей характеристикой рассеяния.

Дисперсией называется средний квадрат отклонения значений признака от среднего арифметического. Дисперсия σ^2, вычисляемая по выборочным данным, называется выборочной дисперсией.

Для несгруппированных данных выборочная дисперсия равна:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ или}$$

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2 \text{ (несмещенная дисперсия);}$$

где $\sum_{i=1}^{n}(x_i - \bar{x})^2$ - сумма квадратов отклонений значений признака x_i от среднего арифметического x.

Для сгруппированных в интервальный вариационный ряд данных:

$$s^2 = \frac{1}{n}\sum_{i=1}^{n} n_i (x_i - \bar{x})^2;$$

где x_i - средние значения интервалов группировки;

$\sum_{i=1}^{n} n_i (x_i - \bar{x})^2$ - взвешенная сумма квадратов отклонений.

Стандартным или **средним квадратическим отклонением** называется корень квадратный из дисперсии: $s = \sqrt{s^2}$.

Размерность стандартного отклонения в отличие от размерности дисперсии совпадает с единицами измерения варьирующего признака, поэтому в практической статистике для того, чтобы охарактеризовать рассеяние признака используют обычно стандартное отклонение, а не дисперсию.

Коэффициент вариации - это отношение среднего квадратического отклонения к среднему значению:

$$k_{вар} = s / \bar{x}.$$

Он измеряет разброс в относительных единицах, в то время как среднее квадратическое отклонение – в абсолютных.

Коэффициент вариации используется и как показатель однородности выборочных наблюдений. Считается, что если коэффициент вариации не превышает 10 %, то выборку можно считать однородной, т. е. полученной из одной генеральной совокупности.

К **показателям формы распределения** экспериментальных данных относятся коэффициенты асимметрии и эксцесса, которые вычисляются на основе моментов третьего и четвертого порядка соответственно.

Коэффициент асимметрии характеризует "скошенность" распределения относительно симметричного нормального распределения. Он является безразмерной величиной и определяется как: $ass = \mu_3 / s^3$;

где $\mu_3 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^3$ - выборочный центральный момент третьего порядка для несгруппированных данных и $\mu_3 = \frac{1}{n}\sum_{i=1}^{n} (n_i x_i - \bar{x})^3$ для данных, сгруппированных в интервальный вариационный ряд.

Если ass = 0, то распределение симметрично, если ass > 0, то преобладают положительные отклонения от математического ожидания,

если ass < 0 - отрицательные. Например, распределение на рисунке 4б) имеет отрицательную асимметрию.

Коэффициент эксцесса характеризует островершинность распределения относительно нормального распределения, для которого эксцесс равен 3:

$$exc = (\mu_4 / s^4) - 3;$$

где $\mu_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4$ - выборочный центральный момент четвертого порядка для несгруппированных данных и $\mu_4 = \frac{1}{n}\sum_{i=1}^{n}(n_i x_i - \bar{x})^4$ для сгруппированных.

Если exc > 0, то распределение имеет более острый пик по сравнению с нормальным распределением, если exc < 0, то распределение имеет плосковершинную форму по сравнению с нормальным распределением.

Все вышеописанные числовые характеристики могут быть рассчитаны для данных, измеренных в шкале отношений.

Для результатов измерений в порядковой шкале при небольшом числе градаций (различных значений) единственным информативным показателем описательной статистики является гистограмма.

Если число градаций велико, то информативными могут быть также мода и медиана данных, сгруппированных по возрастанию признака.

3. ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

3.1. Общая методика проверки гипотез

Статистическая гипотеза представляет собой некоторое предположение о законе распределения случайной величины или о параметрах этого закона, формулируемое на основе выборки.

Примерами статистических гипотез являются предположения: генеральная совокупность распределена по экспоненциальному закону; математические ожидания двух экспоненциально распределенных выборок равны друг другу. В первой из них высказано предположение о виде закона распределения, а во второй – о параметрах двух распределений.

Гипотезы, в основе которых нет никаких допущений о конкретном виде закона распределения, называют **непараметрическими**, в

противном случае – **параметрическими**.

Гипотезу, утверждающую, что **различие между сравниваемыми характеристиками отсутствует**, а наблюдаемые отклонения объясняются лишь случайными колебаниями в выборках, на основании которых производится сравнение, называют **нулевой** или основной гипотезой H_0. Наряду с основной гипотезой рассматривают и **альтернативную**, противоречащую ей гипотезу H_1.

Проверка гипотезы основывается на вычислении некоторой случайной величины – **критерия Z**, точное или приближенное распределение которого известно. Значение критерия является функцией от элементов выборки $Z = Z(x_1, x_2, \ldots, x_n)$.

Наиболее часто используются критерии, приводящие либо к нормальному распределению, либо к распределению χ^2, либо к Т - распределению Стьюдента, либо к распределению Фишера.

Множество возможных значений случайной величины Z можно разбить на два непересекающихся подмножества, как показано на рисунке 8:

- критическая область (1,3) содержит значения критерия, при которых нулевая гипотеза отклоняется,
- область принятия гипотезы (2) – значения Z при которых она принимается.

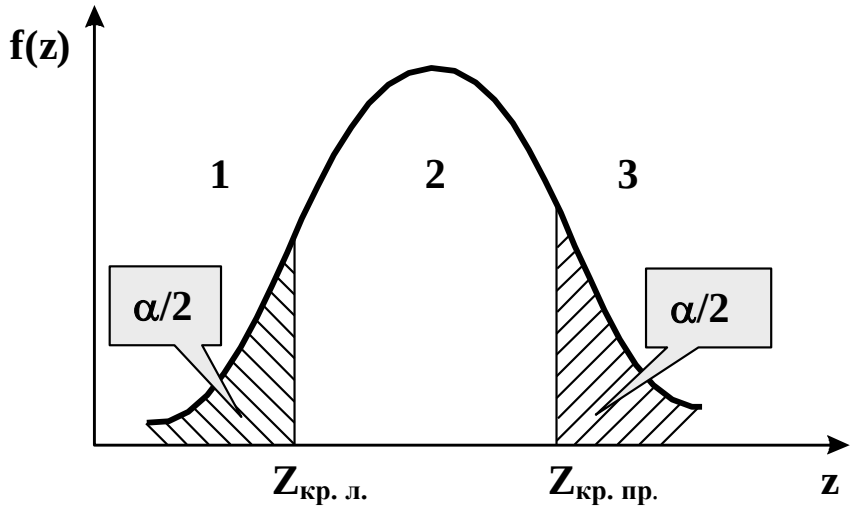


Рисунок 8. Двухсторонняя критическая область критерия.

Значения Z, отделяющие критическую область от области принятия гипотезы H_0, называются **критическими точками** $Z_{кр}$. Критическая область может быть правосторонней, если она находится

правее $Z_{кр.пр}$, левосторонней, если она находится левее $Z_{кр.л}$, или двусторонней, когда $Z < Z_{кр.л}$ и $Z > Z_{кр.пр}$.

Если в процессе проверки нулевая гипотеза будет отвергнута, то верна альтернативная гипотеза.

При проверке гипотезы могут быть допущены ошибки двух видов (рисунок 9):

- **ошибка первого рода**, если отклонена верная нулевая гипотеза,
- **ошибка второго рода**, если принята неверная нулевая гипотеза.

Для нахождения критических точек нужно задать уровень значимости. **Уровнем значимости** α называется вероятность ошибки первого рода, заключающейся в отклонении (не принятии) нулевой гипотезы, когда она верна, то есть вероятность того, что различия сочтены существенными, а они на самом деле случайны. Тогда, например, правосторонняя критическая область задается условием $p(Z > Z_{кр.пр}) = \alpha$.

Ошибка второго рода возникает с вероятностью β в том случае, когда принимается неверная гипотеза H_0, в то время как справедлива конкурирующая гипотеза H_1.

Вероятность отвергнуть ложную гипотезу H_0 называется **мощностью критерия.**

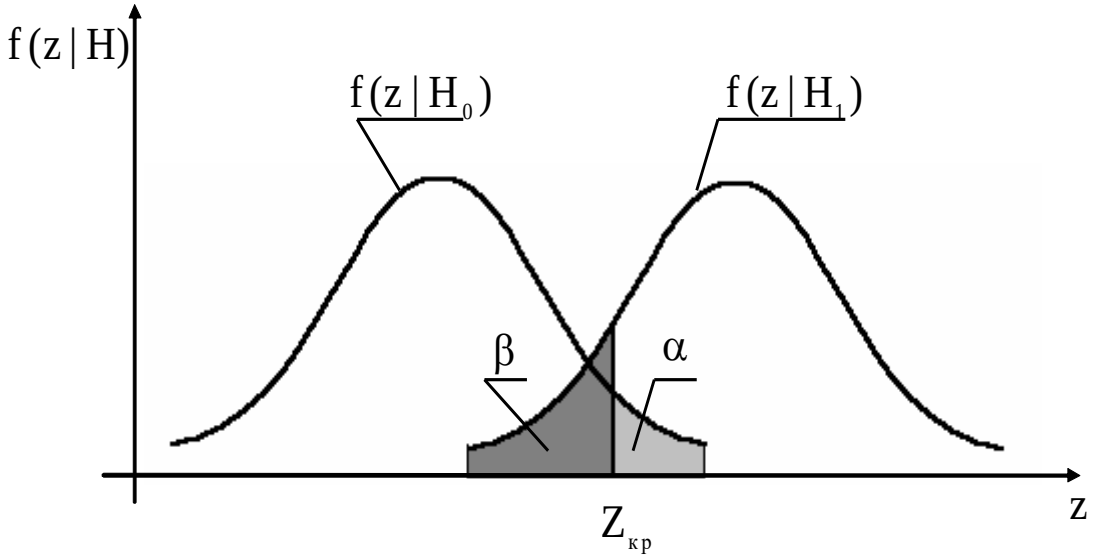


Рисунок 9. Ошибки первого и второго рода при проверке гипотезы

При заданном объеме выборки вероятность совершения ошибки первого рода можно уменьшить, снижая уровень значимости α. Однако при этом увеличивается вероятность ошибки второго рода β (снижается

мощность критерия). Единственный способ уменьшить обе вероятности состоит в увеличении объема выборки

Уровень значимости назначается в зависимости от допустимой вероятности ошибки эксперимента. Обычно используют уровни значимости равные 0,05, 0,01 и 0,001.

В медико-биологических экспериментальных исследованиях ограничиваются значением 0,05, то есть допускается не более чем 5%-ая вероятность ошибки.

Общая **методика проверки** статистических гипотез заключается в следующем.

Этап 1. Выдвигают предположение - нулевую (основную) гипотезу H_0.

Этап 2. Задают величину уровня значимости α.

Этап 3. Для гипотезы H_0 выбирается статистический критерий (решающее правило), как функция от результатов наблюдений $Z(x_1, x_2, ..., x_n)$.

Этап 4. Вычисляют или находят с использованием таблиц квантилей соответствующих распределений критические точки $Z_{кр.}$ в зависимости от уровня значимости α, числа опытов n, вида функции Z.

Этап 5. В функцию Z подставляют имеющиеся конкретные выборочные данные и подсчитывают эмпирическое значение $Z_{эмп}$.

Если **эмпирическое значение критерия** оказывается **меньше или равно критическому**, то **принимается нулевая гипотеза** – считается, что на заданном уровне значимости (то есть при том значении α, для которого рассчитано критическое значение критерия), сравниваемые характеристики совпадают.

В противном случае, если **эмпирическое значение критерия** оказывается **строго больше критического**, то **нулевая гипотеза отвергается** и принимается альтернативная гипотеза – сравниваемые характеристики считаются различными с достоверностью α.

Другими словами, чем меньше эмпирическое значение критерия по сравнению с критическим, тем больше степень совпадения характеристик сравниваемых объектов. И наоборот, чем больше эмпирическое значение критерия, тем сильнее различаются характеристики сравниваемых объектов.

Данные, полученные в реальных экспериментах, могут быть представлены независимыми или сопряженными выборками. К этим выборкам применяются критерии значимости для независимых или для

сопряженных выборок соответственно.

Критерии **для независимых выборок** применяются, чтобы выявить статистическую значимость различий двух разных исследуемых групп. Например:

- параметры двух групп пациентов, к которым применялись различные методики лечения, с целью изучения значимости различий между методиками;

- параметры двух групп пациентов, экспериментальной, к которой применялось воздействие методики, и контрольной, к которой не применялось, с целью изучения значимости данной методики на результаты лечения.

Критерии, применяемые к выборкам с попарно сопряженными вариантами, называются парными критериями или критериями **для связанных** (сопряженных) **выборок**. Например:

- параметры одной и той же испытуемой группы до и после воздействия;

- параметры одного и того же объекта, но относящиеся к различным его частям.

Критерии разделяются на две большие группы: параметрические и непараметрические.

Особенностью **параметрических критериев** является предположение о том, что распределение признака в генеральной совокупности, из которой взята исследуемая выборка, подчиняется нормальному закону распределения. Нормальность эмпирического распределения выборки должна быть доказана до применения параметрического критерия.

Непараметрические критерии (критерии, свободные от распределения) используются для статистической проверки гипотез, когда закон распределения исходной генеральной совокупности неизвестен.

Параметрические критерии являются более мощными, чем их непараметрические аналоги. Если существует возможность использования параметрических критериев, но применяются непараметрические, увеличивается вероятность ошибки второго рода, т.е. принятия ложной нулевой гипотезы.

4.2. Проверка гипотез о законе распределения. Критерий хи-квадрат.

Проверка гипотез о законе распределения осуществляется с помощью **критериев согласия**.

Задача заключается в проверке соответствия эмпирического и гипотетического законов распределения. Нулевая гипотеза утверждает, что различие между эмпирическим и теоретическим законами (например нормальным) значимо и, следовательно, рассматриваемую случайную величину с большой вероятностью нельзя считать нормально распределенной. Альтернативная же предполагает отсутствие значимых отличий, а значит и согласованности эмпирического и теоретического распределений.

Таким образом, критерий согласия, как и другие статистические критерии, должен подтвердить или отвергнуть нулевую гипотезу.

Задача проверки соответствия эмпирических распределений гипотетическим очень важна. Выдвигая гипотезу о согласии эмпирического распределения известному теоретическому, исследователь фактически выбирает статистическую модель исследуемого процесса, которую он будет использовать при его анализе. Если, например, критерий покажет, что закон распределения, построенный по наблюдаемым значениям исследуемой величины, согласуется с нормальным, то можно считать, что она является нормально распределенной. Это очень важно при применении статистических методов анализа, поскольку во многих из них используется предположение о нормальности распределения исследуемых величин

В обработке данных применяется большое количество различных критериев согласия. Среди них наиболее популярными являются критерии хи-квадрат Пирсона, Колмогорова-Смирнова, и др.

Простейшим методом проверки соответствия эмпирического распределения теоретическому является глазомерный метод. Для этого на гистограмму накладывается кривая плотности вероятности предполагаемого теоретического распределения.

Для грубой проверки нормальности распределения можно использовать коэффициенты асимметрии и эксцесса:

- распределение считается симметричным, если $|A| < 0,1$, и сильно асимметричным при $|A| > 0,5$;

- распределение считается близким к нормальному, если $|E| < 0,1$, и сильно отклоняющимся от нормального при $|E| > 0,5$.

Наиболее универсальным из критериев согласия является **критерий согласия Пирсона или критерий χ^2.** Он позволяет проверять

гипотезы о различных законах распределения.

Если эмпирическое распределение задано в виде в виде соседних интервалов (Δx_i) и соответствующих им частот (n_i), то **эмпирическое значение критерия Пирсона** в этом случае вычисляется по формуле:

$$\chi^2_{\text{эмп}} = \sum_{i=1}^{m} \frac{(n_i - n_i^T)^2}{n_i^T},$$

где n_i^T - теоретические частоты для проверяемого закона распределения.

Если эмпирическое распределение задано в виде последовательности x_i, то предварительно необходимо:

- расположить последовательность по возрастанию;
- разбить на интервалы, например, по формуле Стерджесса;
- подсчитать частоту попадания x_i в каждый интервал.

Критическое значение $\chi^2_{\text{кр}}$ находится по таблице квантилей χ^2-распределения по заданному уровню значимости α и числу степеней свободы $\nu = m - 1 - k$, где k - количество параметров закона распределения , оцениваемых по выборке.

Теоретические частоты n_i^T для **нормального закона распределения** можно получить с использованием функции Лапласа по соотношению:

$$n_i^T = n \cdot P_i^T = n\left[\Phi\left(\frac{x_{i+1} - \tilde{m}_x}{\tilde{\sigma}_x} \right) - \Phi\left(\frac{x_i - \tilde{m}_x}{\tilde{\sigma}_x} \right) \right],$$

где x_i, x_{i+1} - границы *i*-го интервала в эмпирическом распределении (в гистограмме);

$\tilde{m}_x, \tilde{\sigma}_x$ - выборочные оценки математического ожидания и среднего квадратического отклонения случайной величины X;

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

При этом $\nu = m - 3$.

Для **показательного закона:**

$$n_i^T = n \cdot P(x_i < X < x_{i+1}) = n(e^{-\lambda x_i} - e^{-\lambda x_{i+1}});\ \nu = m - 2.$$

Для **биномиального распределения:**

$$n_i^T = n \cdot P_i^T = n \cdot C_m^{x_i} P^{x_i} (1-P)^{m-x_i};\ \nu = m - 2,$$

где P - вероятность появления события в каждом испытании; x_i - число появлений события в одном опыте, состоящем из m независимых

испытаний $(x_i = 0, 1, 2, \ldots, m-1)$.

Для **закона Пуассона:** $n_i^T = n \cdot \frac{a^{x_i}}{x_i!} e^{-a}; \ \nu = m - 2.$

Ограничения при использовании критерия Фишера:

- число членов выборки должно быть не менее 30;
- в каждом интервале должно быть не менее 5 членов выборки.

3.3. Параметрические критерии проверки гипотез о параметрах распределения.

Параметрические критерии предназначены для определения достоверности совпадений и различий параметров распределения исследуемых объектов.

Гипотезы о достоверности совпадений и различий средних значений чаще всего проверяется по критерию Стьюдента и его модификациям.

Критерий Стьюдента или **t-критерий** - общее название для серии статистических критериев, в которых статистика критерия имеет распределение Стьюдента. Наиболее часто t-критерии применяются для проверки равенства средних значений в двух выборках.

Нулевая гипотеза предполагает, что средние равны. Отрицание этого предположения называют гипотезой сдвига H_1.

Все разновидности критерия Стьюдента являются параметрическими и основаны на дополнительном предположении о нормальности выборки данных. Поэтому перед применением критерия Стьюдента требуется выполнить проверку нормальности.

Рассмотрим следующие варианты использования критерия Стьюдента для **независимых** выборок.

Вариант 1: **генеральные совокупности X и Y распределены нормально, причем известны их дисперсии D(X) и D(Y).**

Из генеральных совокупностей извлечены две независимые выборки объемом соответственно n и m. Причем, n и m должны быть больше 30.

При заданном уровне значимости α проверяется нулевая гипотеза о равенстве математических ожиданий генеральных совокупностей:

H_o: M (X) = M (Y).

Статистическим критерием для проверки этой гипотезы является

случайная величина:

$$Z = \frac{M(X) - M(Y)}{\sqrt{\frac{D(X)}{m} + \frac{D(Y)}{n}}},$$

имеющая стандартное нормальное распределение.

Наблюдаемое значение критерия:

$$Z_н = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{D(X)}{m} + \frac{D(Y)}{n}}}.$$

Вид критической области зависит от типа конкурирующей гипотезы:

- H_1: M (X) ≠ M (Y), т.е. критическая область двусторонняя, $Z_{кр}$ определяется как аргумент функции Лапласа, при котором

$$\Phi(Z_{кр}) = \frac{1-\alpha}{2},$$

и критическая область задается неравенством $|Z| > Z_{кр}$.

- H_1: M (X)>M (Y), т.е. критическая область правосторонняя, $Z_{кр}$ определяется как аргумент функции Лапласа, при котором

$\Phi(Z_{кр}) = \frac{1-2\alpha}{2}$, и критическая область определяется неравенством $Z > Z_{кр}$.

- H_1: M (X) < M (Y), т.е. критическая область левосторонняя, заданная неравенством $Z < -Z_{кр}$, где $Z_{кр}$ вычисляется так же, как в предыдущем случае.

Вариант 2: **генеральные совокупности X и Y распределены нормально, дисперсии D(X) и D(Y) известны и равны между собой.**

Если известные дисперсии генеральной совокупности равны между собой, выражение для Z-критерия упрощается:

$$Z = \frac{M(X) - M(Y)}{\sqrt{\frac{1}{m} + \frac{1}{n}}}.$$

Вариант 3: **генеральные совокупности X и Y распределены нормально, дисперсии D(X) и D(Y) равны между собой, но неизвестны.**

Из генеральных совокупностей извлечены две независимые выборки объемом соответственно n и m. Причем, n и m меньше 30

(малые выборки).

В качестве критерия для проверки нулевой гипотезы

H_o: M (X) = M (Y) служит случайная величина:

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

имеющая при справедливости нулевой гипотезы распределение Стьюдента с $\nu = n + m - 2$ степенями свободы.

Наблюдаемое значение критерия вычисляется по формуле:

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

Критическая область также строится в зависимости от вида конкурирующей гипотезы:

- H_1: M (X) ≠ M (Y), критическая область двусторонняя, задаваемая неравенством: $|T| > t_{двуст.кр.}$,

где $t_{двуст.кр.}(\alpha,\nu)$ находится из таблицы квантилей распределения Стьюдента.

- H_1: M (X) > M (Y) – критическая область правосторонняя, определяемая условием $T > t_{прав.кр.}$.

- H_1: M (X) < M (Y) – критическая область левосторонняя,

определяемая условием $T < - t_{прав.кр.}$.

Критическая точка вновь находится по таблице квантилей распределения Стьюдента.

Вариант 4: **генеральные совокупности X и Y распределены нормально, дисперсии D(X) и D(Y) не равны между собой и неизвестны.**

Проверка гипотезы о генеральных средних двух независимых групп с нормальным распределением и неизвестными неравными дисперсиями в математической статистике называется проблемой Беренса-Фишера и имеет в настоящее время только приближенные решения.

Если выборки **зависимы,** т.е. взяты из одной генеральной совокупности с неизвестной дисперсией D(X) и имеют одинаковый объем n, то сравнение средних возможно по **критерию Стьюдента для зависимых выборок**:

$$T_н = \frac{\sqrt{n-1}\sum_{i=1}^{n}(x_i - y_i)}{\sqrt{n \cdot \sum_{i=1}^{n}(x_i - y_i)^2 - (\sum_{i=1}^{n}(x_i - y_i))^2}},$$

где $(x_i - y_i)$ - разности одноименных элементов зависимых выборок.

В этом случае проверяется нулевая гипотеза H_o: $\bar{x} = \bar{y}$ при альтернативной гипотезе H_1: $\bar{x} \neq \bar{y}$.

Таким образом, для использования критерия Стьюдента при проверке равенства двух средних генеральных совокупностей по независимым выборкам необходимо, чтобы дисперсии этих совокупностей были известны, либо неизвестны, но равны между собой.

Для сравнения двух дисперсий нормальных генеральных совокупностей чаще всего применяется **критерий Фишера** или **F-критерий**.

Пусть имеются две выборки объемов n_1 и n_2, извлеченные из нормально распределенных генеральных совокупностей X и Y. Требуется по исправленным выборочным дисперсиям s_x^2 и s_y^2 проверить нулевую гипотезу о равенстве генеральных дисперсий рассматриваемых генеральных совокупностей: H_o: D (X) = D (Y).

Критерием служит случайная величина $F = \frac{s_б^2}{s_м^2}$ – отношение большей несмещенной дисперсии к меньшей.

Если $F_н < F_{кр}$ нулевая гипотеза принимается, если $F_н > F_{кр}$ нулевую гипотезу отвергают.

Этот критерий при условии справедливости нулевой гипотезы имеет распределение Фишера-Снедекора со степенями свободы $\nu_1 = n_1 - 1$ и $\nu_2 = n_2 - 1$.

Критическая область зависит от вида конкурирующей гипотезы.

Если H_1: D(X) > D(Y), то критическая область правосторонняя: $p(F > F_{кр}(\alpha, \nu_1, \nu_2)) = \alpha$.

При конкурирующей гипотезе H_1: D(X) ≠ D(Y) критическая область двусторонняя: $p(F < F_1) = \frac{\alpha}{2}$, $p(F > F_2) = \frac{\alpha}{2}$.

Критические точки $F_{кр}(\alpha, \nu_1, \nu_2)$ или $F_{кр}(\frac{\alpha}{2}, \nu_1, \nu_2)$ находятся по таблице квантилей распределения Фишера.

Иногда в статистических расчетах приходится иметь дело с **несколькими выборками**, относительно которых надо решить вопрос об однородности их эмпирических дисперсий. Другими словами, надо решить вопрос, в одинаковых ли условиях, с одинаковой ли погрешностью получены выборки и, следовательно, можно ли сравнивать их между собой.

Для проверки **гипотезы об однородности** эмпирических **дисперсий** нескольких выборок следует пользоваться **критерием Кохрена**, который основан на законе распределения отношения максимальной эмпирической дисперсии $\{s_i^2\}_{max}$ к сумме всех дисперсий, то есть

$$G = \frac{\{s_i^2\}_{max}}{\sum_{i=1}^{n} s_i^2}.$$

Это распределение имеет степени свободы $\nu_1 = n - 1$, где n - объем одной выборки и $\nu_2 = k$, где k – число выборок и меняется в пределах $0 < G < 1$.

Важным условием применения критерия Кохрена является **одинаковый объем** n во всех k выборках.

Затем по выбранному уровню значимости α и степеням свободы ν_1 и ν_2 по таблице критических значений критерия Кохрана определяется $G_{кр}$.

Если вычисленное значение $G < G_{кр}$, то принимается нулевая гипотеза H_0: $s^2_1 = s^2_2 = \ldots = s^2_i = \ldots = s^2_n$ об однородности всех эмпирических дисперсий.

Если выполняются соотношение $G > G_{кр}$, то нулевая гипотеза отвергается и принимается альтернативная гипотеза H_1: $s^2_1 = s^2_2 = \ldots = s^2_i = \ldots = s^2_{max}$, то есть, максимальная дисперсия существенно отличается от остальных.

4.4. Непараметрические критерии проверки гипотез

Непараметрические критерии предназначены для определения достоверности совпадений и различий экспериментальных данных, когда закон распределения генеральной совокупности неизвестен или отличается от нормального.

Критерии, которые рекомендуется применять при непараметрической проверке гипотез для данных, измеренных в различных шкалах, приведены в таблице .

Таблица .

Шкала измерений	Статистический критерий
Отношений	Колмогорова-Смирнова, Крамера-Уэлча, Вилкоксона-Манна-Уитни
Порядковая	Вилкоксона-Манна-Уитни, χ^2 Фишера
Номинальная	Угловое преобразование Фишера

Двухвыборочный критерий однородности Колмогорова-Смирнова позволяет обнаружить расхождения в форме двух эмпирических распределений.

С помощью этого критерия проверяется гипотеза H_0: $F_э(x) = F_э(y)$ о том, что функции распределения $F_э(x)$ и $F_э(y)$ совпадают против альтернативной гипотезы H_1: $F_э(x) \neq F_э(y)$ о том, что они различны.

Критерий Колмогорова-Смирнова $D_{m,n}$ определяется как максимум модуля разности между эмпирической функцией $F_э(x)$, построенной по выборке $x_1, x_2, \ldots x_n$, и эмпирической функцией $F_э(x)$, построенной по выборке $y_1, y_2, \ldots y_m$:

$$D_{m,n} = \max_x \left| F_э(x) - F_э(y) \right|.$$

При справедливости гипотезы H_0 статистика

$$\lambda = D_{m,n} \sqrt{\frac{mn}{m+n}} < \lambda_{кр}$$

имеет асимптотическое распределение Колмогорова, а $\lambda_{кр}$ определяется из условия $P\{\lambda \succ \lambda_\alpha\} = \alpha$, где α- уровень значимости.

Критические значения критерия Колмогорова приведены в сокращенной таблице .

Таблица . Критические значения критерия Колмогорова-Смирнова.

α	0,1	0,05	0,01
λ_α	1,22	1,36	1,63

Алгоритм проверки гипотезы H_0 заключается в следующем:

- строятся эмпирические функции распределений X и Y;

- определяется максимум модуля разности между эмпирическими функциями распределений;

- рассчитывается экспериментальное значение статистики λ;

- из сравнения λ и $\lambda_{кр}$ делается вывод о принятии или отклонении нулевой гипотезы.

Критерий Вилкоксона-Манна-Уитни применяется для проверки сходства или различия двух выборок, измеренных во всех шкалах, кроме номинальной.

Проверка по этому критерию может осуществляться не только по абсолютным значениям элементов, но и по результатам их парных сравнений.

Условия применения критерия:

- сравниваемые выборки имеют эмпирические распределения одинаковой формы и различаются расположением;

- выборки должны быть независимы;

- количество элементов в каждой выборке должно быть не менее 3.

Однако, так как правило принятия решений основано на асимптотической нормальности статистики критерия, то объемы выборок должны быть не менее 10. Иначе надо пользоваться специальными таблицами.

Методика расчета критерия Вилкоксона-Манна-Уитни заключается в следующем.

Имеются две сравниваемые выборки:

$X = (x_1, x_2 \ldots, x_n)$,

$Y = (y_1, y_2 \ldots, y_m)$.

Для каждого элемента первой выборки x_i определяем число a_i элементов второй выборки, которые превосходят его по своему значению (то есть, число таких y_j, что $y_j > x_i$), а также число b_i элементов второй выборки, которые по своему значению равны ему (то есть число таких y_j, что $y_j = x_i$).

Сумма

$$a_1 + a_2 + \ldots + a_n + \frac{1}{2}(b_1 + b_2 + \ldots + b_n) = \sum_{i=1}^{n} a_i + \frac{1}{2}\sum_{i=1}^{n} b_i$$

по всем n членам первой выборки называется эмпирическим значением критерия Манна-Уитни и обозначается U.

Выражение для определения эмпирического значения критерия Вилкоксона:

$$W_н = \frac{\left|\frac{nm}{2} - U\right|}{\sqrt{\frac{nm(n+m+1)}{12}}}.$$

Алгоритм определения достоверности совпадений и различий для экспериментальных данных, измеренных в шкале отношений или интервалов, с помощью критерия Вилкоксона-Манна-Уитни такой же, как и в предыдущих случаях:

- вычислить для сравниваемых выборок $W_н$ - эмпирическое значение критерия Вилкоксона по формуле;

- сравнить $W_н$ с критическим значением $W_{0,05} = 1,96$.

Если $W_н \leq 1,96$, то характеристики сравниваемых выборок совпадают на уровне значимости 0,05; если $W_н > 1,96$, то достоверность различий характеристик сравниваемых выборок составляет 95%.

Методика определения достоверности совпадений и различий с помощью **критерия χ^2-Фишера** для экспериментальных данных, измеренных в порядковой шкале, заключается в следующем.

Характеристикой группы является число ее членов, набравших тот или иной балл.

Для каждой группы формируется вектор баллов :

- для экспериментальной группы $N = (N_1, N_2, \ldots, N_L)$,

- для контрольной группы $M = (M_1, M_2, \ldots, M_L)$,

где L – число баллов порядковой шкалы.

Эмпирическое значение критерия однородности χ^2 для порядковых шкал вычисляется по следующей формуле:

$$\chi^2_н = nm\sum_{i=1}^{L}\frac{\left(\frac{N_i}{n} - \frac{M_i}{m}\right)^2}{N_i + M_i},$$

где n и m – общее количество членов сравниваемых выборок,

N_i и M_i - количество членов сравниваемых выборок, набравших i-тый балл.

Критические значения критерия $\chi^2_{кр}$ выбираются по таблицам квантилей распределения Фишера в зависимости от числа степеней свободы и уровня значимости.

Число степеней свободы для данных, измеренных в порядковой шкале равно L – 1.

Использование критерия χ^2–Фишера для сравнения корректно, только тогда, когда в любой из сравниваемых выборок не менее пяти ее членов получили данный балл.

Поэтому альтернативой может служить переход к дихотомической шкале.

Методика определения достоверности совпадений и различий для экспериментальных данных, измеренных в дихотомической шкале, заключается в следующем.

Для экспериментальной группы, описываемой двумя числами $N_1 + N_2 = n$, доля p ее членов, набравших максимальный балл, равна: $p = N_2 / n$.

Для контрольной группы, $M_1 + M_2 = m$, доля q ее членов, набравших максимальный балл, равна: $q = M_2 / m$.

Для данных, измеренных в дихотомической шкале целесообразно использование в качестве критерия **угловое преобразование Фишера**:

$$\varphi_н = \left|2\arcsin(\sqrt{p}) - 2\arcsin(\sqrt{q})\right|\sqrt{\frac{mn}{m+n}}.$$

Проверка значимости критерия осуществляется путём нахождения вероятности полученного значения в таблицах критических значений t-распределении Стьюдента в зависимости от количества степеней свободы $\nu = (n_1+n_2-2)$ и уровня значимости.

При использовании углового преобразования Фишера на объем выборок накладываются следующие ограничения.

Нижний предел - 2 наблюдения в одной из выборок. При этом:

- если в одной выборке ровно 2 наблюдения, то в другой должно быть не менее 30;
- если в одной выборке ровно 3 наблюдения, во второй должно быть не менее 7;
- если в одной выборке ровно 4 наблюдения, во второй должно быть не менее 5.

5. ИССЛЕДОВАНИЕ ЗАВИСИМОСТЕЙ

5.1 Корреляционный анализ данных

Во многих задачах обработки результатов эксперимента требуется установить взаимосвязь между переменными величинами на основе выборочных данных.

Переменные могут быть независимы, связаны функциональной или статистической зависимостью.

Связь между двумя переменными x и y называется **функциональной**, если определенному значению переменной x строго соответствует одно или несколько значений другой переменной y.

Если обе переменные или одна из них подвержены действию случайных факторов, среди которых есть факторы, общие для обеих переменных, между ними возникает статистическая зависимость.

Статистической называется зависимость, при которой изменение одной из случайных величин x приводит к изменению **распределения** другой случайной величины y.

Если при изменении одной величины изменяется **среднее значение** другой, то статистическая зависимость называется **корреляционной**.

Задачей **корреляционного анализа** является измерение тесноты связи между переменными.

При изучении корреляционной зависимости между переменными возникают следующие задачи:

- измерение силы (тесноты) связи;
- построение корреляционной модели;
- проверка значимости параметров связи.

Для измерения тесноты связи в зависимости от количества элементов в анализируемых выборках, от типа шкалы, в которой производились измерения исходных данных, от требуемой достоверности результатов сравнения и т.д. могут использоваться различные показатели. Основные из них следующие.

Коэффициенты ассоциации и контингенции.

Они применяются для определения тесноты связи двух качественных признаков, каждый из которых состоит только из двух групп.

Для их вычисления строится таблица, которая показывает связь между двумя явлениями, каждое из которых должно быть **альтернативным**, то есть состоящим из двух качественно отличных друг от друга значений признака (например, изделие годное или бракованное).

Пусть провели n наблюдений за двумя признаками А и Б ($n = n_1 + n_2 + n_3 + n_4$) и результаты занесли в таблицу, которая называется четырехклеточной **таблицей сопряженности:**

	Признак А	
Признак Б	да	нет

да	n_1	n_2
нет	n_3	n_4

Тогда коэффициент ассоциации K_a определяется по выражению:

$K_a = \frac{n_1 n_4 - n_2 n_3}{n_1 n_4 + n_2 n_3}$, причем $-1 \le K_a \le +1$.

Пример 26.

Было опрошено 500 человек по двум показателям: наличие или отсутствие у них прививки против гриппа и факт заболевания или незаболевания гриппом во время его эпидемии.

Таблица 15.

Группа лиц	Число лиц	
	заболевших гриппом	не заболевших гриппом
сделавших прививку	30	30
не сделавших прививку	120	120

Данные опроса сведены в таблицу 15, по которой рассчитан коэффициент ассоциации:

$$K_a = \frac{n_1 n_4 - n_2 n_3}{n_1 n_4 + n_2 n_3} = \frac{30 \cdot 80 - 120 \cdot 270}{30 \cdot 80 + 120 \cdot 270} = -0{,}86.$$

Существенный недостаток коэффициента ассоциации: если в одной из четырех клеток таблицы сопряженности частота равна 0, то $|K_a| = 1$, и тем самым преувеличена мера действительной связи.

Чтобы этого избежать, К. Пирсон предложил другой показатель - **коэффициент контингенции**:

$$K_k = \frac{n_1 n_4 - n_2 n_3}{\sqrt{(n_1 + n_2)(n_3 + n_4)(n_1 + n_3)(n_2 + n_4)}}.$$

Для примера 26 коэффициент контингенции равен:

$$K_k = \frac{30 \cdot 80 - 270 \cdot 120}{\sqrt{300 \cdot 200 \cdot 150 \cdot 350}} = -0{,}534.$$

Связь считается достаточно значительной и подтвержденной, если $|K_a| > 0{,}5$ или $|K_k| > 0{,}3$.

Поэтому в примере 26 оба коэффициента характеризуют достаточно большую обратную зависимость между исследуемыми

признаками. Таким образом, можно предположить, что прививка положительно влияет на предупреждение заболевания.

Линейный коэффициент корреляции.

Линейный коэффициент корреляции r_{xy} или коэффициент корреляции Пирсона выражает степень тесноты линейной связи между двумя переменными, измеренными в количественных шкалах, а если форма связи между переменными еще не определена, его рассчитывают с целью получить ответ на вопрос, можно ли считать зависимость линейной.

По выборочным данным линейный коэффициент корреляции вычисляется по следующей формуле:

$$r_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \bar{x})^2 \cdot \sum_{i=1}^{n}(y_i - \bar{y})^2}}.$$

Числитель формуле расчета коэффициента корреляции, деленный на n, т.е.

$$\frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n} = \overline{(x_i - \bar{x})(y_i - \bar{y})},$$

представляет собой среднее произведения отклонений значений двух признаков от их средних, называемое **выборочным корреляционным моментом** или **ковариацией**. Поэтому можно сказать, что линейный коэффициент корреляции представляет собой частное от деления ковариации между x и y на произведение их средних квадратических отклонений:

$$r_{xy} = \frac{\overline{xy} - \bar{x}\bar{y}}{s_x s_y}.$$

В отличие от ковариации коэффициент корреляции является безразмерной величиной, поэтому его можно использовать для сопоставления разных статистических рядов. Например, корреляция между ростом и весом будет одной и той же, независимо от того, проводились измерения в дюймах и фунтах или в сантиметрах и килограммах.

Основные свойства коэффициента корреляции:

- коэффициент корреляции характеризует степень статистической связи между переменными, но не устанавливает причинно-следственную связь между ними, т.е. $r_{xy} = r_{yx}$;

- умножение переменных x и y на постоянные коэффициенты или сложение их с некоторыми постоянными величинами не изменяет коэффициент корреляции: $r_{xy} = r(ax + b, cy + d)$;

- коэффициент корреляции находится в диапазоне $-1 \le r_{xy} \le 1$.

Графическое представление экспериментальных данных виде точек x_i и y_i на плоскости в системе координат xy называется **диаграммой рассеяния** или **корреляционным полем**.

Как показано на рисунке 14, если $\overline{xy} > \bar{x}\bar{y}$, то r_{xy} **Error! Reference source not found.**будет положительным, что характеризует прямую связь между x и y, в противном случае ($r_{xy} < 0$) – обратную связь. Если $\overline{xy} = \bar{x}\bar{y}$, то $r_{xy} = 0$, что означает отсутствие линейной зависимости между x и y, а при $r_{xy} = 1$ зависимость между x и y становится функциональной.

Таким образом, коэффициент корреляции при линейной зависимости служит как мерой тесноты связи, так и показателем, характеризующим степень приближения корреляционной зависимости между x и y к линейной.

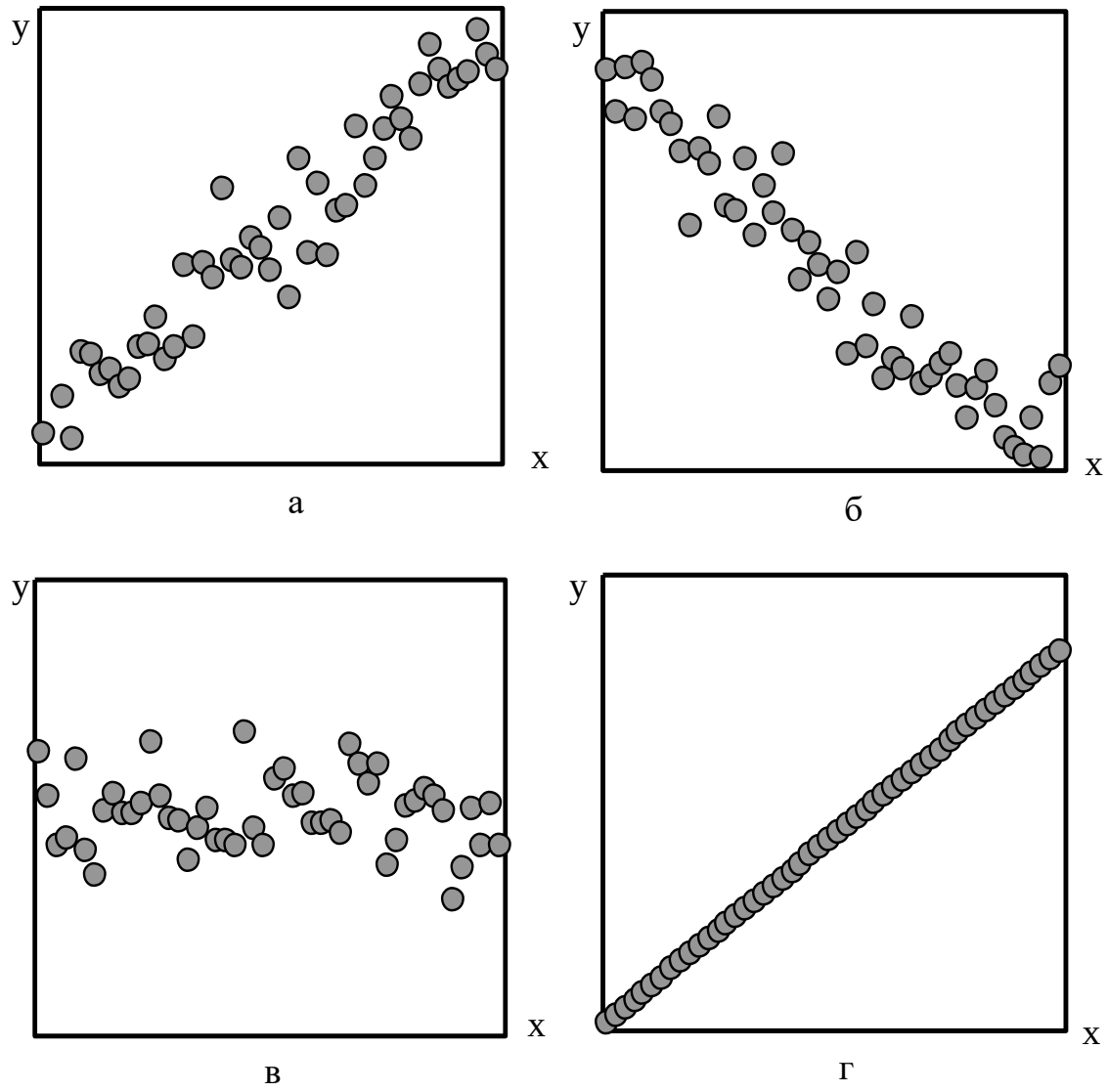


Рисунок 14. Корреляционное поле данных в зависимости от коэффициента корреляции: $1 < r_{xy} > 0$ (а), $-1 > r_{xy} < 0$ (б), $r_{xy} = 0$ (в), $r_{xy} = 1$ (г).

Только по величине коэффициента корреляции нельзя судить о **достоверности** корреляционной связи между признаками. Этот параметр зависит также от числа степеней свободы: $\nu = n - 2$,

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 ($\mathbf{n \leq 30}$), при оценке зависимости между показателями используется следующая градация:

- высокая степень взаимосвязи, когда значения коэффициента корреляции находится в пределах от 0,7 до 0,99;
- средняя степень взаимосвязи, когда значения коэффициента корреляции находятся в пределах от 0,5 до 0,69;
- слабая степень взаимосвязи – значения коэффициента корреляции находятся от 0,2 до 0,49.

Корреляция бисериально - точечная.

Бисериально-точечная корреляция это вид корреляции, которая имеет место в случае, когда один из двух взаимосвязанных признаков выражен в альтернативной форме.

Если имеются две альтернативные группы показателя x численности n_1 и n_2 ($n_1 + n_2 = n$), то коэффициент бисериально - точечной корреляции находится путем определения групповых средних ($\bar{x}_1, \bar{x}_2$) и общего среднего квадратического отклонения σ :

$$K_\sigma = \frac{\bar{x}_1 - \bar{x}_2}{\sigma} \sqrt{\frac{n_1 n_2}{n(n-1)}}.$$

Коэффициент детерминации есть квадрат коэффициента корреляции случайных величин x и y: $K_д = (r_{xy})^2$.

Коэффициент детерминации представляет собой долю вариации, общую для двух переменных

Частные (парциальные) коэффициенты корреляции.

Частные коэффициенты корреляции используются для оценки тесноты связи между двумя показателями из нескольких при исключенном влиянии других показателей.

Например, установлена прямая корреляционная связь между смертностью в результате эпидемий гриппа и количеством врачей, участвующих в ликвидации эпидемий.

Причина такой корреляции в том, что имеется третья переменная (начальный размер эпидемии), которая влияет как на число погибших, так и на число врачей. Если "контролировать" эту переменную (например, рассматривать только эпидемии определенной величины), то исходная корреляция либо исчезнет, либо, возможно, даже изменит свой знак.

Пусть имеется три показателя x, y, v. Частный коэффициент корреляции между x и y при исключении v определяется через парные коэффициенты корреляции соотношением:

$$r_{y|x,v} = \sqrt{\frac{r_{yx} - r_{yv}r_{xv}}{(1-r_{yv}^2)(1-r_{xv}^2)}}.$$

Ранговые коэффициенты Спирмена и Кендалла оценивают степень тесноты связи между двумя ранговыми (качественными или порядковыми) показателями.

Пусть имеем n объектов, которые характеризуются двумя качественными показателями A и B.

Проранжируем объекты в порядке ухудшения качества по показателю A и присвоим объектам ранги x_i, равные их порядковому номеру в этом ряду, т.е. $x_i = i$. Затем при данном расположении объектов припишем ранг y_i по признаку B. Тогда ранговый **коэффициент корреляции Спирмена** вычисляется по формуле

$$r_c = 1 - \frac{6}{n^3 - n}\sum_{i=1}^{n}(x_i - y_i)^2.$$

Допустим, что справа от y_1 имеется k_1 рангов больше чем y_1, а справа от y_2 - k_2 рангов больше чем y_2, ... , справа от y_{n-1} - k_{n-1} рангов больше, чем y_{n-1}.

Тогда ранговый **коэффициент корреляции Кендалла** вычисляется по формуле

$$r_k = \frac{4}{n(n-1)}\sum_{i=1}^{n-1}k_i - 1.$$

Оба коэффициента по модулю меньше единицы и при больших n между значениями r_c и r_k наблюдается определенное соотношение $r_k/r_c \approx 2/3$.

Преимущества ранговых коэффициентов корреляции Спирмена и Кендалла: они легко вычисляются, с их помощью можно изучать и измерять связь не только между количественными, но и между качественными атрибутивными признаками, ранжированными определенным образом. Кроме того, при использовании ранговых коэффициентов корреляции не требуется знать форму связи изучаемых явлений.

Однако мощность их меньше, чем мощность коэффициента корреляции Пирсона.

После определения показателей корреляции следующим этапом корреляционного анализа является **определение значимости** предполагаемой **статистической связи** между случайными переменными x и y.

Для данных, измеренных в количественных шкалах, понятие значимости основывается на предположении, что распределение остатков (т.е. отклонений наблюдений от среднего) для переменной y является нормальным с постоянной дисперсией для всех значений переменной x.

Перед определением значимости корреляций целесообразно визуально оценить распределение экспериментальных данных в пространстве координат xy. При этом следует учитывать следующие особенности распределения переменных.

Выбросы.

Выбросы являются нетипичными, резко выделяющимися наблюдениями. Они могут существенно повлиять на значение коэффициента корреляции.

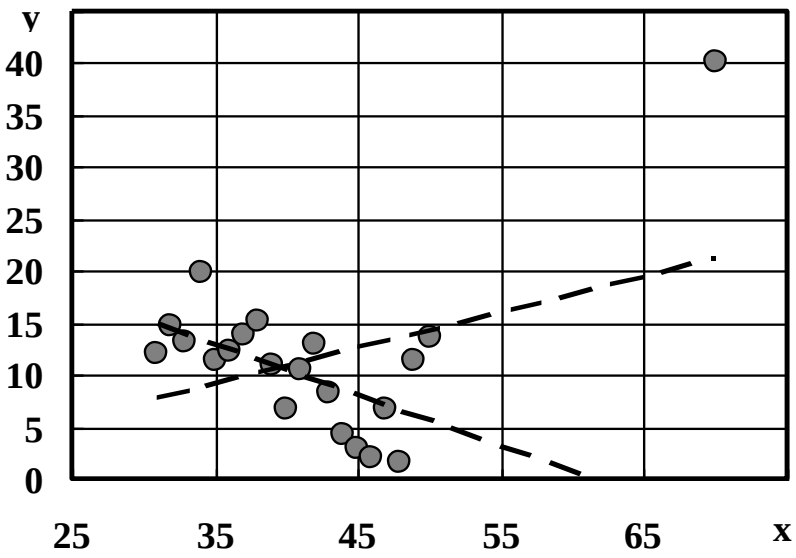


Рисунок 16. Корреляционное поле экспериментальных данных с выбросом.

Например, как показано на рисунке 16, если анализе взаимосвязи между x и y учитывать выброс, то r_{xy} = 0,36, если выброс не учитывать то r_{xy} = -0,6.

Если размер выборки относительно мал, то добавление или исключение некоторых данных, которые, возможно, не являются выбросами, способно оказать существенное влияние на коэффициент корреляции. Увеличение количества членов выборки повышает значимость корреляции.

Общепринятого метода автоматического удаления выбросов не существует. Иногда применяются численные методы удаления. Например, исключаются значения, которые выходят за границы ±2 стандартных отклонений вокруг выборочного среднего. При этом необходимо обязательно перепроверить экспериментальные данные в зоне выбросов и при необходимости скорректировать схему эксперимента.

Корреляции в неоднородных группах.

Отсутствие однородности в выборке также является фактором, смещающим (в ту или иную сторону) выборочную корреляцию.

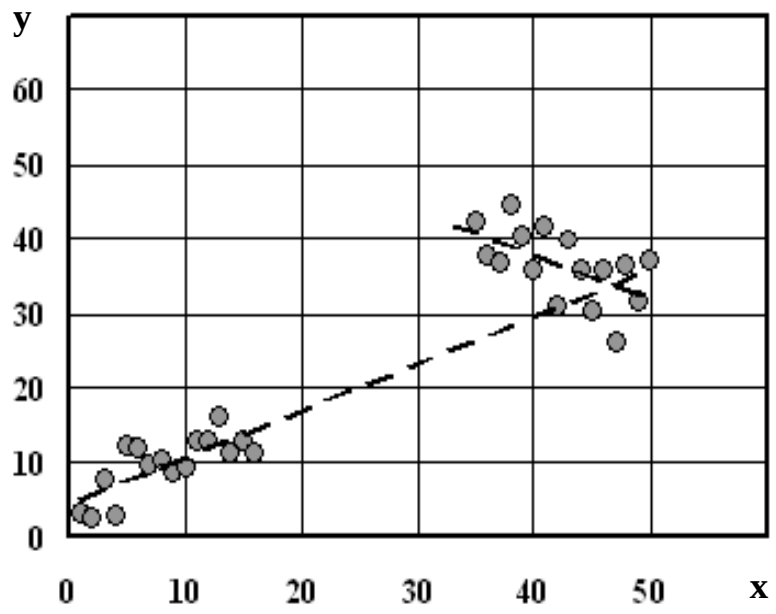


Рис. 17. Корреляционное поле корреляции в неоднородных группах.

Если коэффициент корреляции вычислен по данным, которые поступили из двух различных экспериментальных групп, необходима

обязательная предварительная проверка этих групп на однородность.

Нелинейные зависимости между переменными.

Отклонения от линейности также сильно влияют на коэффициент корреляции. Из рисунка 18 видно, что на нем представлена явно функциональная нелинейная зависимость. Однако r_{xy} = -0,06.

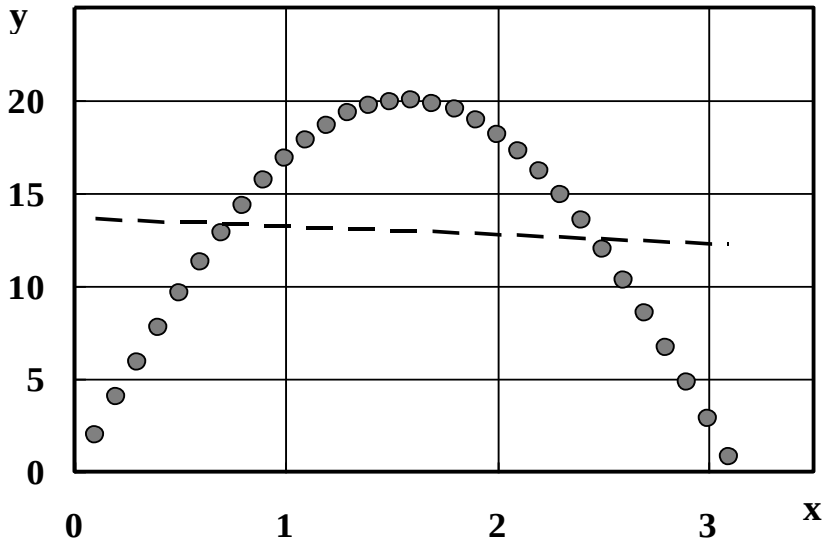


Рисунок 18. Корреляционное поле при нелинейной связи между переменными.

Если корреляция сильная, однако, зависимость явно нелинейная, существует несколько способов описания таких данных:

- можно преобразовать одну или обе переменные, чтобы сделать зависимость линейной, а затем уже вычислить корреляцию между преобразованными величинами, для чего часто используется логарифмическое преобразование;

- можно использовать непараметрическую корреляцию;

- можно попытаться найти функцию, которая наилучшим способом описывает данные, и проверить ее степень согласия с данными.

Последний способ наиболее сложный, но его можно реализовать с помощью соответствующего программного обеспечения.

Количественная оценка значимости корреляции заключается в выдвижении и проверке соответствующих статистических гипотез. Причем, проверяется нулевая **гипотеза о статистической независимости** случайных переменных x и y.

Проверка гипотезы о значимости выборочного парного **линейного коэффициента корреляции** ($H_0 : r_{xy} = 0$) осуществляется с использованием статистики $T_н$, имеющей распределение Стьюдента:

$$T_н = r_{xy} \cdot \sqrt{\frac{n-2}{1-r_{xy}^2}}, \qquad T_{кр} = \mp T(1-\frac{\alpha}{2}; n-2).$$

Проверка гипотезы о значимости выборочного коэффициента **ранговой корреляции Спирмена** ($H_0 : r_{xy} = 0$) осуществляется аналогично с использованием статистики $T_н$, имеющей распределение Стьюдента:

$$T_н = r_{xy} \cdot \sqrt{\frac{n-2}{1-r_{xy}^2}}, \qquad T_{кр} = \mp T(1-\frac{\alpha}{2};\ n-2).$$

Проверка гипотезы о значимости выборочного коэффициента **ранговой корреляции Кендалла** ($H_0 : r_k = 0$) осуществляется с использованием статистики $Z_н$, имеющей нормальное распределение:

$$Z_н = r_k \cdot \sqrt{\frac{9n(n+1)}{2(2n+5)}}.$$

Сравнение двух коэффициентов корреляции ($H_0 :\ r_1 = r_2$) осуществляется с использованием критерия, приводящего к нормальному закону распределения:

$$Z_н = \frac{z_1 - z_2}{\sqrt{\sigma_{z_1}^2 + \sigma_{z_2}^2}},$$

где $\quad z_i = \frac{i}{2} \ln \frac{1+r_i}{1-r_i}, \qquad \sigma_{z_i}^2 = \frac{1}{n_i - 3}, \quad i = 1,2.$

Проверка значимости **теоретического корреляционного отношения** η_T ($H_0 :\ \eta_T = 0$) осуществляется по критерию Фишера:

$$F_н = \frac{\eta_T^2 (n-m)}{(1-\eta_T^2)(m-1)}, \qquad F_{кр} = F(1-\alpha,\ m-1,\ n-m),$$

где n - число опытов, m - число интервалов (групп) различных значений y.

5.2. Регрессионный анализ данных.

Понятия регрессии и корреляции непосредственно связаны между собой, но при этом существует четкое различие между ними.

Если корреляционный анализ дает ответ на вопрос, какова степень связи между переменными, то **регрессионный анализ** - это метод установления аналитического выражения зависимости между этими переменными.

В регрессионном анализе данные состоят из пар значений **зависимой** или **результативной** переменной и **независимой** или **факторной** переменной.

В принципе не имеет значения, какая из сравниваемых переменных будет назначена зависимой, а какая - независимой т.к. регрессионный, как и корреляционный, анализ позволяет определять количественную меру связи переменных, но ничего не говорит о причинно-следственной зависимости.

При проведении так называемого **пассивного эксперимента** и факторная переменная $x\{x_1, x_2, \dots x_n\}$, и результативная переменная $x\{y_1, y_2, \dots y_n\}$ рассматриваются как случайные величины. Однако при проведении **активного эксперимента** экспериментатор может выбирать значения независимой переменной. Тогда независимая переменная может рассматриваться как неслучайная.

В общем случае зависимая переменная есть сумма значений некоторой функциональной модели и случайной величины:

$y = f(x) + \varepsilon$,

где $f(x)$ - функция регрессионной зависимости,

ε - аддитивная случайная величина с нулевым математическим ожиданием.

Предположение о характере распределения этой величины называется гипотезой порождения данных. Обычно предполагается, что величина ε имеет нормальное распределение с нулевым средним и дисперсией σ^2.

Регрессионный анализ решает следующие задачи:

- выбор формы связи между переменными, т.е. модели регрессии;
- оценка неизвестных параметров модели;
- проверка соответствующих статистических гипотез о регрессии.

Форма связи между переменными, т.е. вид уравнения регрессии, выбирается исследователем либо из каких-то теоретических предпосылок, либо из соображений удобства работы с этой формулой, либо из вида и анализа графического изображения имеющихся статистических данных, либо из совокупности возможных математических моделей, либо из других каких-либо предпочтений.

Наиболее универсальным является класс полиномиальных функций:

$y = a^0 + a^1x + a^2x^2 + \ldots + \varepsilon$.

Для такого класса задача выбора функции сводится к задаче выбора значений коэффициентов a_0, a_1, a_2, …, a_n, … . Однако универсальность полиномиального представления обеспечивается только при возможности неограниченного увеличения степени полинома, что не всегда допустимо на практике, поэтому приходится применять и другие виды функций. Например:

- показательную функцию $y = ab^x$;
- дробно-рациональную функцию $y = (ax + b)^{-1}$;
- логарифмическую функцию $y = a \cdot \ln(x) + b$;
- гиперболическую функцию $y = a + b/x$;
- дробно-рациональную функцию $y = x/(ax + b)$.

Частным случаем, широко применяемым на практике, является полином первой степени или **уравнение линейной регрессии** y на x: $y = ax + b + \varepsilon$.

В этом уравнении параметр a - **свободный член**, графически он представляет, как показано на рис. 19, отрезок ординаты y.

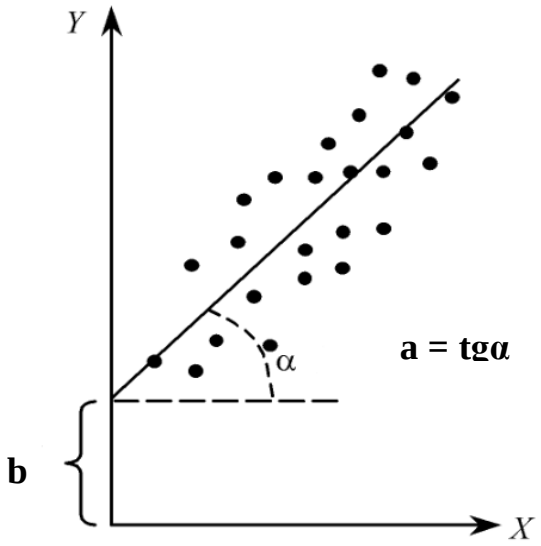


Рис. 19. Графическое изображение уравнения линейной регрессии

Параметр b называется **коэффициентом регрессии** или угловым коэффициентом, определяющим наклон линии регрессии по отношению к осям координат. Этот параметр показывает, насколько в среднем величина зависимой переменной y изменяется при изменении на единицу меры независимой переменной x, то есть вариацию y, приходящуюся на единицу вариации x.

Параметры в уравнении регрессии обычно выбираются по методу наименьших квадратов. Основной принцип этого метода заключается в том, чтобы так определить неизвестные параметры модели, чтобы сумма квадратов отклонений имеющихся данных от выбранной кривой (уравнения) регрессии была бы минимальной: $\sum_{i=1}^{n}(y_i - \bar{y}_{x_i})^2 \to \min$, где $\bar{y}_{x_i} = f(x_i)$.

Если в качестве функциональной модели регрессии выбрана линейная модель, то согласно этому методу необходимо найти минимум функции

$$F = \sum_{i=1}^{N}(y_i - ax_i - b)^2 .$$

Используя условие экстремума функции F, найдем

$$\begin{cases} \dfrac{\partial F}{\partial a} = 2\sum_{i=1}^{n}(y_i - ax_i - b)\cdot x_i = 0 \\ \dfrac{\partial F}{\partial b} = 2\sum_{i=1}^{n}(y_i - ax_i - b) = 0. \end{cases}$$

От последней системы можно перейти к более простой, выполнив над ней элементарные алгебраические преобразования:

$$\begin{cases} \sum_{i=1}^{n} y_i \cdot x_i - a\cdot\sum_{i=1}^{n} x_i \cdot x_i - b\cdot\sum_{i=1}^{n} x_i = 0 \\ \sum_{i=1}^{n} y_i - a\cdot\sum_{i=1}^{n} x_i - b = 0. \end{cases}$$

Решая последнюю систему относительно A и B, получим:

$$a = \frac{\sum_{i=1}^{n} x_i \cdot \sum_{i=1}^{n} y_i - n\cdot\sum_{i=1}^{n} x_i \cdot y_i}{\left(\sum_{i=1}^{n} x_i\right)^2 - n\cdot\sum_{i=1}^{n} x_i^2};$$

$$b = \frac{1}{n}\cdot\left(\sum_{i=1}^{n} y_i - a\cdot\sum_{i=1}^{n} x_i\right).$$

В случае поиска уравнения регрессии в виде полинома k-й степени, исходя из метода наименьших квадратов:

$$Z = \sum_{i=1}^{n}(y_i - \bar{y}_{x_i})^2 = \sum_{i=1}^{n}(y_i - a_0 - a_1 x_i - a_2 x_i^2 - \ldots - a_k x_i^k)^2 \to \min,$$

Вычисляя и приравнивая частные производные критерия Z по каждому неизвестному параметру к нулю ($\frac{\partial Z}{\partial a_j}=0$), получим систему уравнений,

$$\begin{cases} a_0 n + a_1\sum_{i=1}^{n} x_i + a_2\sum_{i=1}^{n} x_i^2 + \ldots + a_k\sum_{i=1}^{n} x_i^k = \sum_{i=1}^{n} y_i, \\ \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad \cdot \\ a_0\sum_{i=1}^{n} x_i^k + a_1\sum_{i=1}^{n} x_i^{k+1} + a_2\sum_{i=1}^{n} x_i^{k+2} + \ldots + a_k\sum_{i=1}^{n} x_i^{2k} = \sum_{i=1}^{n} y_i x_i^k, \end{cases}$$

решая которую, находят неизвестные параметры a_j.

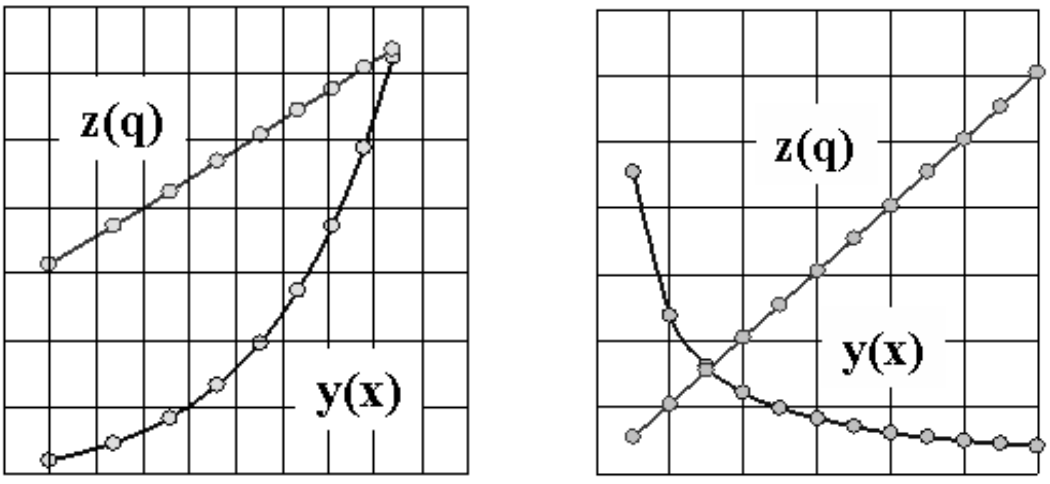
В некоторых случаях, когда визуальный анализ корреляционных полей показывает явный тип нелинейности предполагаемой зависимости, для нахождения уравнения регрессии используют **метод преобразования координат**. Для этого предполагаемая функция заменяется линейной и уравнение регрессии рассматривается в новых координатах.

Например, дробно - рациональная функция $y = (ax + b)^{-1}$ может быть преобразована в линейную путем следующей замены:

$z = 1/y$, $q = x$, тогда, как показано на рис. 21а, $z = aq + b$.

Аналогично (рис. 21б) можно преобразовать в линейную логарифмическую функцию $y = a\ln(x) + b$:

$q = \ln(x)$, $z = y$, $z = aq + b$.



а б

Рис. 21. Преобразование координат для дробно – рациональной (а) и логарифмической функции (б)

После того, как модель построена, то есть найдены ее параметры, необходимо проверить ее адекватность исходным данным, а также полученную точность.

Оценка адекватности принятой регрессионной модели осуществляется несколькими способами, но все они базируется на анализе **регрессионных остатков** ε_i, т.е. разницы между действительными значениями зависимой переменной y_i и расчетными или теоретическими значениями по уравнению регрессии $\hat{y}$: $\varepsilon_i = y_i - \hat{y}_i$.

Адекватность регрессионной модели означает, что:

- остатки должны быть независимыми случайными величинами;
- распределение остатков должно быть нормальным с нулевым средним;
- дисперсия остатков должна быть одинаковой и постоянной.

Визуально оценить распределение остатков можно по графику зависимости остатков ε_i от теоретических значений зависимой переменной $\hat{y}$, как показано на рис. 22.

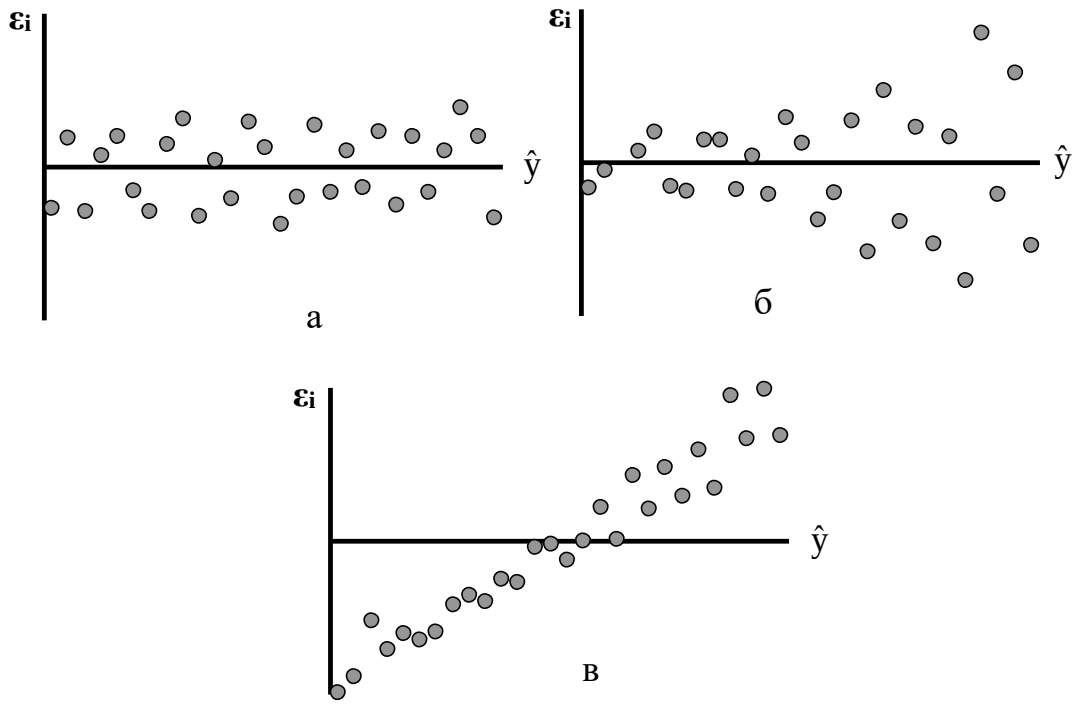


Рис.22. Варианты распределения остатков

Если на графике получена горизонтальная полоса, как показано на рис. 22а, то можно предположить, что остатки ε_i являются случайными величинами, не зависимыми от $\hat{y}$, т.е. линия регрессии хорошо описывает фактические значения y_i.

Если остатки ε_i не случайны, т.е. не имеют постоянной дисперсии (рис. 22б) или зависят от величины $\hat{y}$ (рис. 22в), то в таких случаях необходимо либо применять другую функцию регрессии, либо использовать дополнительную информацию, например, увеличивать количество экспериментальных данных.

Оценка по коэффициенту детерминации.

Статистический смысл коэффициента детерминации R^2 или квадрата коэффициента корреляции заключается в том, что он показывает, какая доля зависимой переменной y объясняется построенной функцией регрессии y(x). Например, при коэффициенте детерминации 0,49 регрессионная модель объясняет 49% дисперсии зависимой переменной, остальные же 51% считаются обусловленными факторами, не отраженными в модели.

Оценка по критерию Фишера.

Оценка значимости уравнения регрессии в целом дается с помощью F-критерия Фишера. В ее основе лежит разложение общей дисперсии зависимой переменной на составляющие.

Общая дисперсия зависимой переменной $S_{общ} = S_ф + S_{ост}$ разлагается на:

- **факторную дисперсию** $S_ф$, обусловленную линией регрессией, которая характеризует воздействие факторной переменной;

- **остаточную дисперсию** $S_{ост}$ относительно линии регрессии, т.е. ту часть вариации зависимой (результативной) переменной, которую нельзя объяснить воздействием независимой (факторной) переменной.

Чем меньше $S_{ост}$, т.е. меньше воздействие неучтенных в модели или случайных факторов, тем точнее соответствует модель фактическим данным. Именно поэтому остаточная дисперсия может быть использована для оценки качества регрессионной модели, точности подбора регрессионной функции.

Основная гипотеза регрессионного анализа H_0: уравнение регрессии не значимо, т.е. факторная и остаточная дисперсии статистически неразличимы.

Для проверки этой гипотезы используется F-критерий:

$F = \dfrac{s_Ф^2 (n-k)}{s_{ОСТ}^2}$, который имеет распределение Фишера с числом степеней свободы $\nu_1 = k - 1$ и $\nu_2 = n - k$,

где n - количество опытов,

k - количество параметров в уравнении регрессии (для линейной модели k = 2),

$s_{\Phi}^2 = \sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2$ - факторная сумма квадратов отклонений,

$s_{ОСТ}^2 = \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$ - остаточная сумма квадратов отклонений.

Суммы квадратов отклонений, отнесенные к числу степеней свободы, представляют собой несмещенные оценки соответствующих дисперсий.

Используя формулу расчета коэффициента корреляции, можно получить следующее выражение F-критерия:

$$F_н = \frac{r_{xy}^2}{1 - r_{xy}^2}(n - 2).$$

Задаваясь уровнем значимости α, и числом степеней свободы ν_1 и ν_2, по таблицам критических значений F-распределения находим $F_{кр} = F(1 - \alpha; k - 1; n - k)$.

Если $F_н > F_{кр}$, H_0 отклоняется, уравнение регрессии признается значимым.

5.4. Дисперсионный анализ данных

Дисперсионный анализ является статистическим методом анализа результатов наблюдений, зависящих от различных одновременно действующих факторов, с целью выбора наиболее значимых факторов и оценки их влияния на исследуемый процесс.

Например, необходимо проанализировать, как влияет степень механической травмы на продолжительность лечения или изучить влияние термической обработки на твердость обрабатываемого материала. В первом случае **фактором** является травма, которая рассматривается на различных **уровнях** (легкая, средняя, тяжелая), во втором случае **фактор** - термическая обработка, **уровни фактора** – закалка, отпуск, нормализация.

Таким образом, дисперсионный анализ применяют, когда хотят выяснить, оказывает или не оказывает влияние на результат эксперимента некоторый **качественный фактор**, который имеет несколько различных **уровней** реализации.

Дисперсионный анализ может быть **однофакторным**, если исследуется влияние на результаты эксперимента одного качественного

фактора, и **многофакторным,** если исследуется влияние нескольких факторов.

Если факторы, чье влияние на величину x исследуется, являются **количественными**, то дисперсионный анализ можно применять и в этом случае. Но более глубокие выводы о характере влияния таких факторов на величину x лучше сделать на основе корреляционно-регрессионного анализа.

Параметрические методы дисперсионного анализа основываются на следующих предпосылках:

- распределение исходных случайных величин нормально;
- дисперсии экспериментальных данных однородны для всех экспериментов, выполненных на различных уровнях изучаемого фактора.

Поэтому при проведении дисперсионного анализа необходимо предварительно проверить **нормальность распределения** результатов экспериментов и **однородность дисперсий** экспериментальных данных.

В основе дисперсионного анализа лежит разделение общей дисперсии на части или компоненты.

Рассмотрим следующий набор данных:

	Группа 1	Группа 2
Наблюдение 1	2	6
Наблюдение 2	3	7
Наблюдение 3	1	5
Групповое среднее	2	6
Групповая сумма квадратов	2	2
Общее среднее	4	
Общая сумма квадратов	28	

Средние двух групп существенно различны (2 и 6 соответственно). Сумма квадратов отклонений внутри каждой группы равна 2. Складывая их, получаем 4.

Если теперь повторить эти вычисления без учета групповой принадлежности, то есть, если вычислить общую сумму квадратов, исходя из общего среднего этих двух выборок, то получим величину 28.

Дисперсия (сумма квадратов), основанная на внутригрупповой изменчивости, приводит к гораздо меньшим значениям, чем при вычислении на основе общей изменчивости (относительно общего среднего). Причина этого, очевидно, заключается в существенной

разнице между средними значениями, и это различие между средними и объясняет существующее различие между суммами квадратов.

Таким образом, общая сумма квадратов (28) разбита на компоненты: сумму квадратов, обусловленную внутригрупповой изменчивостью (2 + 2 = 4) и сумму квадратов, обусловленную различием средних значений между группами (28-(2+2)=24).

В итоге получаем:

$S_{о.} = S_{ф} + S_{ост}$,

$S_{о}$ - общая дисперсия наблюдаемых значений, которая характеризуется разбросом данных от общего среднего и определяет вариацию признака во всей совокупности под влиянием всех факторов, обусловивших эту вариацию;

$S_{ф}$ - межгрупповая или факторная дисперсия, которая характеризуется различием средних на каждом уровне фактора и определяется влиянием исследуемого фактора;

$S_{ост}$ - внутригрупповая или остаточная дисперсия, которая характеризует рассеяние данных внутри уровней фактора и отражает случайную вариацию, т.е. часть вариации, происходящую под влиянием неучтенных случайных факторов и не зависящую от исследуемого фактора.

Таким образом, исследование **компонент дисперсии** служит средством определения значимости **различия между средними**.

Однофакторный параметрический дисперсионный анализ

В однофакторном дисперсионном анализе на одну количественную переменную y оказывает влияние один фактор A (один качественный показатель), наблюдаемый на k уровнях, т. е. имеем k выборок для переменной y.

Например, необходимо проанализировать, как влияет степень и локализация механической травмы на продолжительность лечения или изучить влияние термической обработки и марки стали на ее твердость поле термоупрочнения.

Алгоритм проведения однофакторного дисперсионного анализа состоит в следующем:

- формирование таблицы экспериментальных данных;
- вычисление средних квадратов отклонений;
- вычисление дисперсий;

- оценка результатов анализа.

Исходные данные собираются в следующую таблицу:

Номер наблюдения	Уровни фактора А					
	A_1	A_2	...	A_j	...	A_k
1	y_{11}	y_{21}	...	y_{i1}	...	y_{k1}
2	y_{12}	y_{22}	...	y_{i2}	...	y_{k2}
...	...	...	...	...	...	...
j	y_{1j}	y_{2j}	...	y_{ij}	...	y_{kj}
...	...	...	...	...	...	...
n	y_{1n}	y_{2n}	...	y_{in}	...	y_{kn}

Наблюдаемые данные обозначим y_{ij},

где i-индекс уровня фактора, i = 1,2, ..., k,

j - индекс наблюдения на i-м уровне фактора, j = 1,2, ..., n

На каждом уровне может быть свое число наблюдений n_i.

Тогда общее число опытов (наблюдений) $n = \sum_{i=1}^{k} n_i$.

По данным y_{ij} можно определить следующие характеристики:

- среднее значение переменной y на i-м уровне

$$\bar{y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{ij};$$

- среднее значение переменной y по всем наблюдениям

$$\bar{y} = \frac{1}{n} \sum_{i=1}^{k} \sum_{j=1}^{n_i} y_{ij} = \frac{1}{n} \sum_{i=1}^{k} \bar{y}_i n_i;$$

- сумму квадратов отклонений всех наблюдений от общего среднего

$$s_0 = \sum_{i=1}^{k} \sum_{j=1}^{n_i} (y_{ij} - \bar{y})^2;$$

- сумму квадратов отклонений средних групповых значений от общего среднего $\bar{y}$

$$s_ф = \sum_{i=1}^{k} (\bar{y}_i - \bar{y})^2 n_i;$$

- остаточную сумму квадратов отклонений

$$s_{ост} = \frac{1}{n}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2.$$

Суммы квадратов отклонений, отнесенные к числу степеней свободы, представляют собой несмещенные оценки соответствующих дисперсий.

Далее вычисляются компоненты дисперсии:

- общая дисперсия

$$S_o^2 = \frac{s}{n-1} = \frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y})^2.$$

- факторная (межгрупповая) дисперсия

$$S_ф^2 = \frac{s_ф}{k-1} = \frac{1}{k-1}\sum_{i=1}^{k}(\bar{y}_i - \bar{y})^2 n_i.$$

- остаточная (внутригрупповая) дисперсия

$$S_{ост}^2 = \frac{s_{ост}}{n-1} = \frac{1}{n-k}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(y_{ij} - \bar{y}_i)^2.$$

В параметрическом дисперсионном анализе проверяется гипотеза H_0 о равенстве средних групповых значений количественного показателя ($H_0: \bar{y}_1 = \bar{y}_2 = ... = \bar{y}_k$).

Критерием для проверки этой гипотезы является соотношение межгрупповой дисперсии и внутригрупповой дисперсии: $F_н = \frac{S_ф^2}{S_{ост}^2}$ (F-критерий Фишера).

Однофакторный непараметрический дисперсионный анализ

Непараметрические методы дисперсионного анализа применяются, когда нет возможности проверить нормальность распределения исходных данных, а также когда результаты экспериментов представлены порядковых шкалах.

В качестве непараметрического теста для выявления наличия статистически значимых различий между средними нескольких выборок применяют **критерий Краскела-Уоллиса**.

Он используется для сравнения трех или более выборок, и проверяет нулевые гипотезы, согласно которым различные выборки были взяты из одного и того же распределения, или из распределений с одинаковыми медианами.

Критерий Краскела-Уоллиса использует не количественные значения экспериментальных данных, а их **ранги**, т.е. номера мест, которые занимают эти данные в вариационном ряду выборки.

Алгоритм проверки нулевой гипотезы H_0 состоит в следующем:

- наблюдаемые данные y_{ij} упорядочиваются по возрастанию и заменяются рангами r_{ij};

- для каждого уровня A_j вычисляется средний ранг R_j:

$$R_j = \frac{1}{n_j}\sum_{i=1}^{n} r_{ij};$$

- проверяется гипотеза H_0.

Статистика для проверки нулевой гипотезы имеет вид:

$$H = \frac{12}{n(n+1)}\sum_{i=1}^{k}\frac{R_j^2}{n_j} - 3(n+1),$$

где n – вся совокупность наблюдений.

Критерий H представляет собой дисперсию ранговых сумм.

При $n_i \geq 5$, $k \geq 4$ b справедливости проверяемой гипотезы H_0 статистика подчиняется распределению χ^2 с (k - 1) степенями свободы.

Проверяемая гипотеза отклоняется на уровне значимости α, если при данных реализациях выборок $H > \chi^2_{1-\varepsilon}$, где $\chi^2_{1-\varepsilon}$ - квантиль уровня $(1 - \varepsilon)$ распределения χ^2 с (k - 1) степенями свободы.

Двухфакторный дисперсионный анализ

В двухфакторном дисперсионном анализе на одну количественную переменную y оказывают влияние одновременно два качественных фактора A и B на уровнях A_1, A_2, ...,A_k и B_1, B_2,…,B_m соответственно.

Результаты измерения количественной переменной y представлены двухфакторной таблицей, где:

i = 1,k – число уровней фактора A (число столбцов);

j = 1,n – число уровней фактора B (число строк);

Фактор B	Фактор A						∑
	A_1	A_2	…	A_j	…	A_k	
B_1	y_{11}	y_{21}	…	y_{i1}	…	y_{k1}	Y'_1
B_2	y_{12}	y_{22}	…	y_{i2}	…	y_{k2}	Y'_2
	…	…	…	…	…	…	
B_j	y_{1j}	y_{2j}	…	y_{ij}	…	y_{kj}	Y'_j

	…	…	…	…	…	…	
B_n	y_{1n}	y_{2n}	…	y_{in}	…	y_{kn}	Y'_n
$\sum$	Y_1	Y_2	…	Y_i	…	Y_k	

Дисперсионный анализ для двухфакторных таблиц проводится по следующему алгоритму. Вычисляются суммы:

$$Q_1 = \sum_{i=1}^{k}\sum_{j=1}^{n} y_{ij}^2; \quad Q_2 = \frac{1}{n}\sum_{i=1}^{k} Y_i^2; \quad Q_3 = \frac{1}{k}\sum_{j=1}^{n} Y_j'^2;$$

$$Q_4 = \frac{1}{nk}\left(\sum_{i=1}^{n} Y_j'\right)^2 + \frac{1}{nk}\left(\sum_{i=1}^{k} Y_i\right)^2.$$

Далее находятся оценки дисперсий:

$$S_o^2 = \frac{Q_1 + Q_4 - Q_2 - Q_3}{(k-1)(n-1)};$$

$$S_A^2 = \frac{Q_2 - Q_4}{(k-1)}; \quad S_B^2 = \frac{Q_3 - Q_4}{(n-1)}.$$

Если $\frac{S_A^2}{S_O^2} > F_{кр}(\nu_1, \nu_2)$, где $\nu_1 = n - 1$ и $\nu_2 = (k - 1)(n - 1)$, то влияние фактора A на уровне значимости α признается значимым.

Аналогично значимым признается влияние фактора B, если

$\frac{S_B^2}{S_O^2} > F_{кр}(\nu_1, \nu_2)$, где $\nu_1 = n - 1$ и $\nu_2 = (k - 1)(n - 1)$.

Такой алгоритм предполагает **независимость** факторов A и B. Если они зависимы, то взаимодействие факторов C = AB также является фактором, которому соответствует своя дисперсия. Для того, чтобы выделить такое взаимодействие, необходимы параллельные наблюдения в каждой клетке таблицы, т. е. при каждом сочетании факторов A и B на уровнях A_i и B_j соответственно необходимо не одно наблюдение, а серия наблюдений $y_{ij1}, y_{ij2}, \ldots, y_{ijm}$.

Пусть y_{ij} теперь является средним из m наблюдений, т. е.

$$y_{ij} = \frac{1}{m}\sum_{r=1}^{m} y_{ijr}.$$

Для оценки влияния взаимодействия факторов AB вычисляем дополнительную сумму:

$$Q_5 = \sum_{i=1}^{k}\sum_{j=1}^{n}\sum_{r=1}^{m} y_{ijr}^2.$$

Далее анализ проводится, как и ранее, с той лишь разницей, что в клетках таблицы вместо отдельных значений y_{ijr} используются их средние значения y_{ij}.

Вычисляется дисперсия:

$$S^2_{AB} = \frac{Q_5 - mQ_1}{nk(m-1)},$$

и проверяется значимость взаимодействия факторов АВ критерием:

$$n\frac{S^2_O}{S^2_{AB}} > F_{кр}(\nu_1, \nu_2),$$ где $\nu_1 = (k-1)(n-1)$ и $\nu_2 = nk(m-1)$.

С добавлением каждого нового фактора принципиальная основа дисперсионного анализа не изменяется, но существенно усложняются формулы и таблицы для расчетов.