Учебно-исследовательская работа

Кафедра «Радиоэлектроника»

Лекционный курс

Авторы

Шокова Ю.А.,

Звездина М.Ю.

Аннотация

Лекционный курс предназначен для студентов 2 курса, обучающихся по специальности 11.03.01 Радиотехника.

Авторы

Шокова Юлия Александровна –

КАНДИДАТ ФИЗИКО-МАТЕМАТИЧЕСКИХ НАУК, ДОЦЕНТ КАФЕДРЫ «РАДИОЭЛЕКТРОНИКА»

Звездина Марина Юрьевна –

ДОКТОР ФИЗИКО-МАТЕМАТИЧЕСКИХ НАУК, ДОЦЕНТ, ЗАВ. КАФЕДРОЙ «РАДИОЭЛЕКТРОНИКА»

СОДЕРЖАНИЕ

1 Экспериментальные данные как предмет исследования

1.1 Классификация видов экспериментальных исследований

Исторически сложилось, что познание законов окружающего мира основывалось на обобщении и анализе данных, полученных в результате наблюдения или эксперимента. Накопление знаний об окружающей природе и природных явлениях через эксперимент составляет отдельные экспериментальные разделы наук, например, существуют физика как теоретическая, так и экспериментальная. Ни один теоретический закон не может быть принят, если он противоречит экспериментальным результатам. Рассмотрим, чем наблюдение отличается от эксперимента, и какие виды экспериментов бывают.

Наблюдением называется *способ получения данных, при котором воздействие наблюдателя на объект сведено к минимуму*. ***Экспериментом*** - *наблюдение с воздействием на наблюдаемый объект*.

Экспериментальные исследования в современной науке заканчиваются представлением результатов в виде сформулированных выводов и выдачи рекомендаций. Данная информация может выражаться в виде графиков, чертежей, таблиц, статистических данных или словесных описаний.

Эксперимент предполагает проведение тех или иных опытов. Под ***опытом*** понимают *воспроизведение исследуемого явления в определенных условиях проведения эксперимента при возможности регистрации его результатов*.

По цели проведения и форме представления полученных результатов эксперимент делят на качественный и количественный.

Качественный эксперимент *устанавливает только сам факт существования какого-либо явления, но при этом не дает никаких количественных характеристик объекта исследований*, т.е. качественный эксперимент предусматривает только словесное описание его результатов.

Для анализа свойств объекта в иных условиях, а также формулировке количественных рекомендаций необходимо использовать данные ***количественного эксперимента***, который *не только фиксирует факт существования того или иного явления, но и позволяет установить соотношение между количественными характеристиками явления и количественными характеристиками способов внешнего воздействия на объект исследований*.

Количественный эксперимент предполагает количественное определение всех тех способов внешнего воздействия на объект исследования, от которых зависит его поведение – количественное описание всех факторов.

Фактор – *переменная величина, по предположению влияющая на результаты эксперимента*.

Например, в качестве факторов эксперимента при исследовании удельного сопротивления проводника можно выбрать величину температуры окружающей среды, его качественный состав.

В каждом отдельном опыте каждый из факторов может принимать одно из возможных значений – ***уровень фактора***, т.е. *фиксированное значение фактора относительно начала отсчета*. Фиксированный набор уровней всех факторов в каждом опыте определяет одно из возможных состояний объекта исследований.

При проведении опытов многое зависит от того, насколько активно экспериментатор может вмешиваться в исследуемое явление, имеет ли он или нет возможность устанавливать те уровня факторов, которые представляют для него интерес. С этой точки зрения все факторы можно разбить на три группы:

• ***контролируемые и управляемые*** – это *факторы, для которых можно не только зарегистрировать их уровень, но еще и задать в каждом конкретном опыте любое его возможное значение*;

• ***контролируемые, но неуправляемые факторы*** – это *факторы, уровни которых можно только регистрировать, а задать в каждом опыте их определенное значение практически невозможно*;

• ***неконтролируемые*** – это *факторы, уровни которых не регистрируются экспериментатором и о существовании которых он даже может и не подозревать*.

В количественном эксперименте необходимо не только регистрировать уровни всех контролируемых факторов, но и иметь возможность устанавливать количественное описание того свойства (отклика) исследуемого явления, которое изучает экспериментатор. Поскольку на объект исследования в процессе эксперимента всегда влияет огромное количество неконтролируемых факторов, что вносит в получаемые результаты некоторый элемент неопределенности, значение отклика в каждом конкретном опыте невозможно предсказать заранее. В связи с этим воспроизведение исследуемого явления при одном и том же фиксированном наборе уровней всех контролируемых факторов будет приводить к различным значениям отклика, т.е. отклик – это всегда случайная величины.

Отклик – *наблюдаемая случайная величина, по предположению зависящая от факторов*.

В результате количественного эксперимента необходимо найти зависимость между откликом и факторами – ***функцию отклика***. Данная функция также будет случайной величиной и ее можно задать одним из параметров своего распределения, например, математическим ожиданием.

С учетом приведенного выше деления факторов на три группы функцию отклика можно в самом общем случае записать в виде:

$$M_y = f(x_i, h_i) + \varepsilon_\delta, \qquad (1.1)$$

где x_i - контролируемые и управляемые факторы;

h_i -контролируемые, но неуправляемые факторы;

ε_δ - ошибка эксперимента, учитывающая влияние неконтролируемых факторов.

В зависимости от того, какой группой факторов располагает исследователь, количественный эксперимент можно разделить еще на два вида:

• ***пассивный***, когда в распоряжении исследователя нет управляемых факторов, уровни факторов регистрируются, но не задаются;

• ***активный***, когда экспериментатор имеет возможность не только контролировать факторы, но и управлять ими.

1.2 Типы измерений и измеряемых физических величин

Предметом количественного эксперимента являются количественные величины. Для определения абсолютного значения некоторой физической вели-

чины ее сравнивают с эталоном, который считается единицей величины. Например, единицей длины является метр, времени – секунда, частоты – Герц.

Различают прямое и косвенное измерения. Наиболее простым является ***прямое измерение***, *при котором искомое значение величины находят непосредственно с помощью измерительного прибора*. Например, длина измеряется линейкой, напряжение – вольтметром и т.п.

Если прямые измерения невозможны, то используют ***косвенные измерения***. В них *искомое значение величины находят на основании известной зависимости этой величины от других, допускающих прямое измерение*. Например, электрическое сопротивление резистора определяют по падению на нем напряжения и току через него.

Измерения могут быть выполнены как однократно, так и многократно. ***Однократное измерение*** *дает единственный результат, который принимают за окончательный результат измерения значения искомой величины*. ***Многократное измерение*** *проводят путем повторения однократных измерений одной и той же постоянной физической величины, что приводит к получению набора данных*. Окончательный результат многократного измерения, как правило, находят из набора данных в виде среднего арифметического результатов всех отдельных измерений.

Физические величины, встречающиеся в эксперименте, относят к следующим основным типам:

- случайная величина;
- постоянная величина;
- изменяющаяся (переменная) величина;
- нестабильная величина.

Случайная величина связана со случайными процессами, поэтому результат отдельного измерения данной величины не может быть однозначно предсказан заранее. В то же время проведение достаточно большого числа измерений случайной величины позволяет установить, что результаты измерений отвечают определенным статистически закономерностям. Их выявление, изучение и учет составляют неотъемлемую часть любого эксперимента. В качестве случайных величин можно рассматривать, например, отклонение значения амплитуды сетевого напряжения от номинальной величины.

К ***постоянным величинам*** можно отнести физические постоянные, например, скорость света в вакууме, заряд электрона и т.п. Кроме того, постоянными величинами можно считать некоторые характеристики конкретного объекта, находящегося при фиксированных условиях. Этот тип физических величин чаще всего встречается в экспериментах, например, при определении теплоемкости. Однако многократные измерения постоянной величины могут дать неодинаковые результаты. Дело в том, что результаты измерений подвержены неконтролируемым влияниям многочисленных воздействий внешней среды, включая и неконтролируемые процессы в исследуемых объектах и используемых измерительных приборах. Вследствие этого постоянная величина зачастую проявляет себя как случайная величина, а результат ее измерений отражает случайную природу воздействия и отвечает определенным статистическим закономерностям. В связи с этим для обработки результатов измерения постоянной величины естественно использовать методы, характерные для обработки результатов измерения случайной величины.

Изменяющаяся (переменная) величина закономерно меняется с течением времени вследствие процессов, проходящих в исследуемом объекте.

Примером может служить затухание собственных колебаний в резонаторе. Измерения, проводимые в различные моменты времени, фиксируют величину в новых условиях. Набор результатов однократных измерений представляет собой результаты принципиально неповторимых измерений, так как время нельзя повернуть вспять, а измерение в целом не может расцениваться как многократное.

Особого внимания заслуживает ***нестабильная величина***, которая меняется с течением времени без каких бы то ни было статистических закономерностей. К основной характеристике нестабильной величины следует отнести отсутствие у экспериментатора информации о ее зависимости от времени. Измерения такой величины дают набор данных, не несущих сколько-нибудь полезных сведений. Вместе с тем нестабильная величина может быть переведена в разряд изменяющихся величин, если экспериментально или теоретически установлена закономерность в зависимости ее от времени.

1.3 Типы погрешностей измерений и их оценки

Снятие экспериментальных данных никогда не происходит с абсолютной точностью, а содержит некую ***погрешность***, которая является количественной характеристикой неоднозначности результата измерений. Наличие погрешности обусловлено как принципиально ограниченной точностью измерения, так и природой самих измеряемых объектов. Таким образом, при записи результата измерений возникает необходимость, как оценить значение измеряемой величины, так и указать, насколько полученная оценка близка к истинному значению.

Результат серии многократных измерений может быть рассмотрен как вектор значений случайной величины $X = \{x_1, x_2, \ldots, x_N\}$. Данный вектор является выборкой объема *N* из генеральной совокупности, распределенной по некоторому закону. В качестве оценки значения измеряемой величины, которая является математическим ожиданием в генеральной совокупности, используется среднее значение выборки, которое рассчитывается по формуле:

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i \,. \qquad (1.2)$$

Для характеристики погрешности, с которой была получена оценка результата, обычно используют абсолютную и относительную погрешности.

Абсолютной погрешностью ΔX *называют разность между истинным значением измеряемой величины X и его оценкой - результатом измерения* $\overline{X}$. Данная величина выражается в единицах измеряемой величины *X*. Следует учитывать, то это значение не отражает качества измерений. Например, абсолютная погрешность 1 мм при измерении размеров помещения свидетельствует о высоком качестве измерения. Однако та же погрешность совершенно неприемлема при измерении диаметра тонкой проволоки.

Критерием качества измерения является ***относительная погрешность*** - *отношение абсолютной погрешности к результату измерения*:

$$\varepsilon = \frac{\Delta X}{\overline{X}} \,. \qquad (1.3)$$

Данная величина безразмерна и при расчетах измеряется в долях от единицы. При записи результатов измерения возможно также указывать ее значение в процентах, используя для расчета вместо формулы (1.3) формулу

$$\varepsilon = \frac{\Delta X}{\overline{X}} \cdot 100\% . \qquad (1.4)$$

Высокой точности измерения соответствует малое значение относительной погрешности.

Выделяют следующие ***основные типы ошибок***.

• ***промахи*** - *результаты с аномальным числовым значением* (например, один из результатов измерения отличается от остальных на порядок). Причина промахов - резкое нарушение условий измерения при отдельных наблюдениях, например, сбой аппаратуры, непредвиденное вмешательство и т.д. В многократных измерениях промах появляется обычно не более одного-двух раз; его наличие при обработке результатов измерений может приводить к сильному искажению результата. Отбраковка промахов из результатов измерений является довольно сложным и неоднозначным процессом;

• ***систематические ошибки*** - *постоянная составляющая погрешности, закономерно изменяющаяся при повторных измерениях одной и той же величины*. Систематические ошибки делятся на методологические, возникающие из-за неправильного выбора метода измерения, и инструментальные (приборные). Как правило, при обработке результатов измерений предполагается, что методическая составляющая случайной погрешности максимально учтена введением поправок, и при расчете погрешностей необходимо учитывать только приборную погрешность.

Точность простейших измерительных приборов, например линеек, определяется ценой наименьшего деления. Однако в технических измерениях чаще используются более сложные приборы, которые работают на основе некоторого физического явления и состоят из многих частей. На взаимодействие отдельных частей влияет множество случайных факторов, а точность прибора зависит от допусков на изготовление отдельных частей.

Для многих электроизмерительных приборов погрешность прибора выражается ***классом точности*** *K — приборной погрешностью, выраженной в процентах от максимально допустимых показаний прибора по выбранной шкале измерения* X_m. Существуют как односторонние шкалы (с нулем на конце), так и двусторонние (нуль посередине). Средняя инструментальная погрешность измерительного прибора будет равна

$$\delta = \frac{X_m}{100} K , \quad \text{для односторонней шкалы;} \qquad (1.5)$$

$$\delta = \frac{2X_m}{100} K , \quad \text{для двусторонней шкалы.} \qquad (1.6)$$

При обработке результатов многократных измерений применяется следующий алгоритм действий.

В случае ***прямого многократного измерения*** результаты многократных измерений при одних и тех же условиях $x_1, x_2, \ldots, x_N$ заносятся в таблицу. Далее

1 Проводится оценка истинного значения измеряемой величины по формуле (1.2);

2 Для получения абсолютной погрешности вычисляется оценка среднеквадратического отклонения по формуле:

$$\sigma = \sqrt{\frac{\sum_{i=1}^{N} \left(x_i - \overline{X}\right)^2}{N-1}} , \qquad (1.7)$$

а также вычисляется среднеквадратическая ошибка среднего

$$\sigma_x = \frac{\sigma}{\sqrt{N}}; \quad (1.8)$$

3 При учете инструментальной погрешности в зависимости от типа шкалы прибора вычисляется средняя инструментальная погрешность по формуле (1.5) или (1.6);

4 Априорно задавшись требуемым уровнем доверительной вероятности *P* (чаще всего выбирается значение 0,95), из таблицы распределения Стьюдента находится значение $t(P, N-1)$, а при учете инструментальной погрешности и значение $t(P,\infty)$;

5 Вычисляется абсолютное значение случайной погрешности

$$\Delta X_{сл} = t(P, N-1)\cdot \sigma_x, \quad (1.9)$$

а при учете инструментальной погрешности – и абсолютное значение инструментальной погрешности

$$\Delta X_{пр} = t(P,\infty)\cdot \frac{\delta}{3}; \quad (1.10)$$

6 Абсолютная погрешность прямых измерений рассчитывается с помощью формулы

$$\Delta X = \sqrt{\Delta X_{сл}^2 + \Delta X_{пр}^2} = \sqrt{\left(t(P, N-1)\cdot \sigma_x\right)^2 + \left(t(P,\infty)\cdot \frac{\delta}{3}\right)^2}. \quad (1.11)$$

Если расчет инструментальной погрешности не проводится то $\Delta X = \Delta X_{сл}$.

7 По формуле (1.3) или (1.4) проводится расчет относительной погрешности. Результат измерений записывается в следующем виде

$X = \overline{X} \pm \Delta X$ при доверительной вероятности *P*. Погрешность измерений составляет ε. (1.12)

В случае ***косвенного многократного измерения*** имеется ряд наборов прямых измерений величин *A, B, C,…*, исходя из которых необходимо оценить величину *Z*, связанную с величинами *A, B, C,…* функциональной зависимостью $Z = f(A,B,C,\ldots)$. Поэтому при обработке косвенных измерений сначала необходимо провести обработку результатов прямых измерений величин *A, B, C,…*по приведенному выше алгоритму. После этого

1 Проводится оценка истинного значения измеряемой величины по средним значениям величин *A, B, C,…*

$$\overline{Z} = f\left(\overline{A},\overline{B},\overline{C},\ldots\right); \quad (1.13)$$

2 По средним значениям величин *A, B, C,…* и их значениям абсолютных погрешностей находится абсолютная погрешность искомой величины

$$\Delta Z = \sqrt{\left(\frac{\partial f}{\partial A}\right)^2_{A=\overline{A},B=\overline{B}\ldots}(\Delta A)^2 + \left(\frac{\partial f}{\partial B}\right)^2_{A=\overline{A},\ldots}(\Delta B)^2 + \left(\frac{\partial f}{\partial C}\right)^2_{A=\overline{A},\ldots}(\Delta C)^2 + \ldots}; \quad (1.14)$$

3 На основе оценки $\overline{Z}$ и абсолютной погрешности ΔZ по формуле (1.3) или (1.4) рассчитывается относительная погрешность косвенных измерений.

Для простейших видов функциональной зависимости абсолютная и относительная погрешности могут быть записаны в более простом виде путем преобразования формулы (1.14). Соотношения для некоторых видов функциональных зависимостей приведены в таблице 1.1.

Таблица 1 - Погрешности косвенных измерений для простейших функций

f	Δf	$\varepsilon = \Delta f / f$
A+B	$\sqrt{(\Delta A)^2 + (\Delta B)^2}$	$\frac{\sqrt{(\Delta A)^2 + (\Delta B)^2}}{\overline{A} + \overline{B}}$
A-B	$\sqrt{(\Delta A)^2 + (\Delta B)^2}$	$\frac{\sqrt{(\Delta A)^2 + (\Delta B)^2}}{\overline{A} - \overline{B}}$
A·B	$\sqrt{\overline{B}^2(\Delta A)^2 + \overline{A}^2(\Delta B)^2}$	$\sqrt{\left(\frac{\Delta A}{\overline{A}}\right)^2 + \left(\frac{\Delta B}{\overline{B}}\right)^2}$
$\frac{A}{B}$	$\frac{\sqrt{\overline{B}^2(\Delta A)^2 + \overline{A}^2(\Delta B)^2}}{\overline{B}^2}$	$\sqrt{\left(\frac{\Delta A}{\overline{A}}\right)^2 + \left(\frac{\Delta B}{\overline{B}}\right)^2}$
sin$A$	$\cos\overline{A} \cdot \Delta A$	$\operatorname{ctg}\overline{A} \cdot \Delta A$
cos$A$	$\sin\overline{A} \cdot \Delta A$	$\operatorname{tg}\overline{A} \cdot \Delta A$
tg$A$	$\frac{1}{\cos^2\overline{A}}\Delta A$	$\frac{2}{\sin 2\overline{A}}\Delta A$
ln$A$	$\frac{\Delta A}{\overline{A}}$	$\frac{1}{f}\frac{\Delta A}{\overline{A}}$

Результат косвенных измерений записывается аналогично результату прямых измерений.

2 Характеристика данных выборки и генеральной совокупности

2.1 Принципы подбора выборки

На прошлой лекции было показано, что результатом проведения экспериментальных исследований является некоторая совокупность измерений. При этом однократные измерения допускаются только в виде исключения, так как они не позволяют судить о достоверности измерительной информации, и в основном проводятся многократные измерения.

Данные измерения могут рассматриваться как вектор значений некоторой случайной величины, поскольку исход каждого испытания (результат каждого опыта) подвержен влиянию погрешностей.

Полный набор всех возможных значений, которые может принимать случайная величина при бесконечном числе испытаний, называется ***генеральной совокупностью***. Распределение данных в генеральной совокупности подчиняется некоторым априорно неизвестным законам, которые требуется установить исследователю, причем в его распоряжении никогда нет генеральной совокупности, и он может изучать только ее часть – выборку, причем всегда ограниченного объема.

Выборка *– набор значений величины* $\{x_i\}$*, полученный из генеральной совокупности в результате конечного числа испытаний N. Количество данных в выборке называется ее* ***объемом***. Получить выборку из генеральной совокупности можно разными способами, однако для проведения исследований необходимо, чтобы распределение данных в выборке как можно более точно повторяло характер поведения данных в генеральной совокупности. Сам процесс отбора данных для выборки может являться источником ошибок, *когда характер выборки не полностью воспроизводит характер генеральной совокупности*. Такие ошибки называются ***ошибками репрезентативности***.

Классическим примером влияния ошибок репрезентативности на результат является случай, происшедший во время президентских выборов 1936 года в США. Журнал «Литрери Дайджест», успешно прогнозировавший события нескольких предшествующих выборов, ошибся в своих предсказаниях, разослав десять миллионов пробных бюллетеней своим подписчикам, а также людям, выбранным по телефонным книгам всей страны и людям из регистрационных списков автомобилей. В редакцию вернулось 25% разосланных бюллетеней (почти 2,5 миллиона), в них голоса были распределены следующим образом:

- 57% отдавали предпочтение кандидату-республиканцу Альфу Лэндону
- 40% выбрали действующего в то время президента-демократа Франклина Рузвельта

Однако на выборах, набрав более 60% голосов, победил Рузвельт. Ошибка журнала заключалась в следующем: редакция знала, что большинство подписчиков являлись республиканцами. Желая увеличить репрезентативность выборки, они расширили ее за счёт людей, выбранных из телефонных книг и регистрационных списков. Однако они не учли современных им реалий и набрали ещё больше республиканцев: во время Великой депрессии обладать телефонами и автомобилями могли себе позволить в основном представители среднего и высшего класса (то есть большинство республиканцев, а не демократов).

Правильная организация выборки, ее репрезентативность, позволяет избежать подобных ошибок. Репрезентативность выборки достигается ***рандомизацией*** - *случайным отбором членов из генеральной совокупности*. Это обеспечивает равную возможность для всех членов генеральной совокупности попасть в состав выборки.

При отборе выборки строгое соблюдение правила рандомизации как правило либо труднореализуемо, либо приводит к серьезным затратам на проведение исследования. Тогда в зависимости от задачи применяются различные способы частичной рандомизации.

Статистический анализ полученных данных выборки позволяет:

- дать для больших выборок общие характеристики, отражающие так называемую центральную тенденцию, т.е. число (или ряд чисел), вокруг которых «рассыпаны» данные, а также степень их разброса;
- провести сравнение нескольких выборок и определить вероятность того, что их различия вызваны случайными причинами, а также оценить общие характеристики;
- получить сведения о взаимосвязях элементов в выборке;
- применить результаты анализа для предсказания и описания. Методы предсказания позволяют выяснить, с какой вероятностью может быть верным или неверным сделанный вывод. Описательные методы сообщают информацию о данных без подтверждения какими-либо вероятностными методами.

2.2 Гистограмма и полигон частот

Предварительная обработка данных начинается с определения, какими типами переменных (признаков) представлены данные. Выделяют три *типа переменных*:

- ***непрерывные*** – представлены действительными числами (например, длина или вес);
- ***дискретные*** – представлены целыми, как правило, положительными числами;
- ***категориальные*** (например, марка кабеля, тип материала, географический регион). Значения категориальных данных не могут быть положены на числовую прямую.

Распределение данных в генеральной совокупности, приближением которой можно считать выборку, подчиняется некоторому априорно неизвестному закону - закону распределения. Установление закона распределения данных в генеральной совокупности по экспериментальным данным является одной из центральных задач математической статистики. Более подробно о ней мы поговорим в следующих лекциях, а сейчас рассмотрим, что собой представляет закон распределения данных.

Самым простым способом наглядной характеристики выборочного закона распределения является построение гистограммы или полигона частот.

Рассмотрим ***алгоритм построения гистограммы***.

Пусть выборка представляет собой серию из *N* измерений, содержащую некоторый набор значений $x_1, x_2, \ldots, x_N$. *Гистограмма* – это график, на оси абсцисс O*x* которого откладываются полученные в отдельных измерениях значения x_i, а на оси ординат O*y* – количество значений x_i в выборке. Для более наглядной гистограммы применяется группирование данных.

В первую очередь данные значения сортируют по возрастанию – строят ***вариационный ряд***

$$x_1 \le x_2 \le ... \le x_N . \quad (2.1)$$

Затем вариационный ряд группируется: отрезок [x_1, x_N] разбивается на несколько непересекающихся отрезков – «карманов» - некоторым способом. Выбор числа «карманов» и способа разбиения является неоднозначен.

Два самых известных способа разбиения – на «карманы» равной длины и равновероятностный способ.

При ***разбиении отрезка на «карманы» равной длины*** длина эталонного «кармана» вычисляется исходя из длины исходного отрезка [x_1, x_N] и числа «карманов» по простой математической формуле.

При ***равновероятностном разбиении*** границы «карманов» выбирают так, чтобы в каждом «кармане» было одинаковое число значений (т.е. необходимо, чтобы объем выборки N делился на число «карманов» n без остатка):

$$a_i = \frac{x_{(i-1)\nu} + x_{(i-1)\nu+1}}{2}, \quad (2.2)$$

где

$$\nu = \frac{N}{n}. \quad (2.3)$$

Далее мы будем рассматривать разбиение отрезка на «карманы» равной длины. Выбор числа «карманов» n также неоднозначен. Можно рассчитывать его по эвристической формуле Стерджесса

$$n = 1 + 3{,}322 \cdot \lg N, \quad (2.4)$$

или по формуле Брукса и Каррузера

$$n = 5 \cdot \lg N, \quad (2.5)$$

или через соотношение

$$n = \sqrt{N}. \quad (2.6)$$

При получении дробного значения n следует произвести округление до ближайшего целого в меньшую сторону.

Далее будем применять формулу Стерджесса.

Исходя из геометрических соображений (рисунок 2.1), длина каждого «кармана» будет равна:

$$\Delta = \frac{x_N - x_1}{n}, \quad (2.7)$$

а координаты a_i, b_i каждого интервала [a_i, b_i], i=1…n, будут вычисляться по формулам

$$a_1 = x_1, \quad b_n = x_N, \quad a_i = b_{i-1}, \text{ для } i = 2..n, \quad (2.8)$$

$$a_i = x_1 + (i-1)\Delta, \quad (2.9)$$

$$b_i = x_1 + i\Delta. \quad (2.10)$$

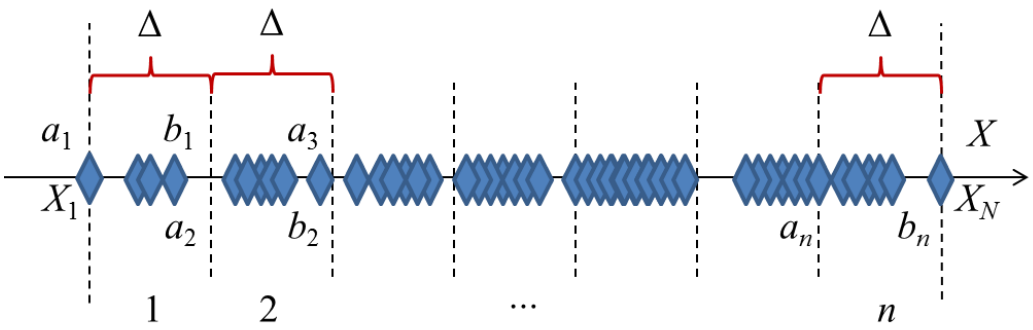


Рисунок 2.1 – Группировка вариационного ряда

После этого необходимо вычислить количество экспериментальных значений, попавших в каждый интервал:

$$T_i = \sum_{j=1}^{N} t_{j,i} \,, \qquad (2.11)$$

где T_i – количество экспериментальных точек в i-м интервале;

$t_{j,i}$ – признак наличия j-й точки в i-том интервале:

$$t_{j,i} = \begin{cases} 1, \text{ если } x_j \in [a_i, b_i], \\ 0, \text{ если } x_j \notin [a_i, b_i]. \end{cases}$$

Значения откладываются по оси ординат, Oy, и строится столбчатый график: ширина столбиков равна длине «кармана», высота i-го столбика – числу T_i. Для практических целей (например, облегчения сравнения гистограмм для двух выборок), гистограмму нормируют, то есть высоту столбца рассчитывают как

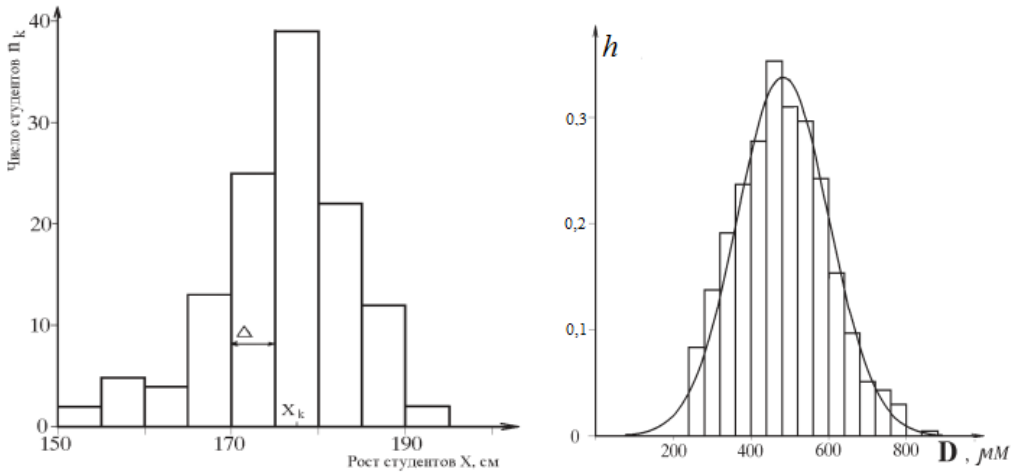
$$h_i = \frac{T_i}{N \cdot \Delta} \,. \qquad (2.12)$$

Пример гистограммы и нормированной гистограммы приведен на рисунке 2.2 (а и б соответственно).

Для нормированной гистограммы произведение высоты h_i на длину «кармана» имеет смысл вероятности попадания результата отдельно измерения в данный интервал. Суммарная площадь под всей гистограммой равна 1.

$$\sum_{i=1}^{N} h_i \Delta = 1 \,. \qquad (2.13)$$

2.4 Если вместо столбиков на графике по середине каждого кармана откладывается вверх величина T_i или h_i, а далее полученные точки соединяются ломаной, то говорят, что был построен полигон частот (для T_i) или полигон относительных частот (для h_i). Примеры полигона частот приведены на рисунке 2.3.



а б

Рисунок 2.2 – Построение гистограммы

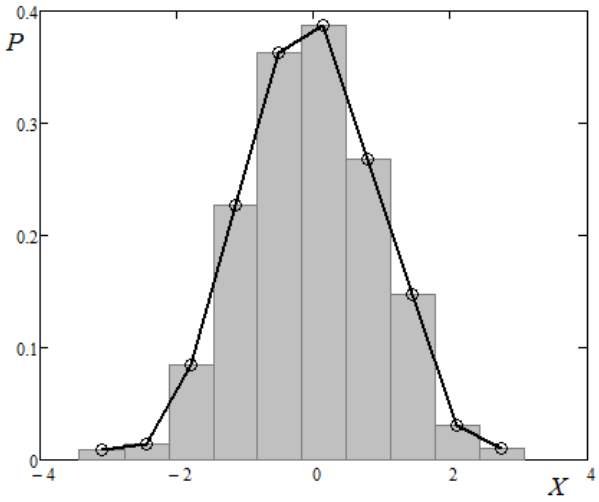


Рисунок 2.3 – Пример построения полигона частот

2.5 Если число измерений *N* достаточно велико ($N \to \infty$), то длину «кармана» можно сделать очень малой ($\Delta \to 0$). Тогда в пределе вместо нормированной гистограммы или полигона относительных частот получим *график, характеризирующий вероятностное распределение значений на генеральной совокупности* – ***кривую распределения***. Функция *f(x)*, описывающая форму кривой распределения, называется ***плотностью вероятности***. Полагается, что

$$\int_{-\infty}^{+\infty} f(x)dx = 1. \qquad (2.14)$$

В случае дискретной величины вместо $f(x)dx$ используют вероятность p_i получить в результате измерения значение x_i.

Вероятность попадания измеряемой величины в интервал $(-\infty, x]$ называют ***функцией распределения*** или ***интегральной функцией распределения***:

$$F(x) = \int_{-\infty}^{x} f(z)dz. \qquad (2.15)$$

Исходя из смысла функции распределения, имеют место следующие соотношения:

$$F(-\infty)=0,\ F(+\infty)=1. \tag{2.16}$$

Первое значение следует понимать как вероятность наступления невозможного события: того, что случайная величина будет меньше -∞. Такая вероятность равна 0. Второе значение описывает вероятность наступления достоверного события: того, что случайная величина будет меньше +∞. Такая вероятность равна 1.

Таким образом, плотность вероятности $f(x)$ равна производной функции распределения $F(x)$ в точке x. Если проинтегрировать плотность вероятности в пределах от x_1 до x_2, то полученная величина будет представлять вероятность *P* того, что результат отдельного измерения будет лежать в интервале $[x_1, x_2]$:

$$\mathrm{P}(x_1 < x < x_2) \equiv \int_{x_1}^{x_2} f(x)dx = F(x_2) - F(x_1). \tag{2.17}$$

2.3 Параметры распределения и их влияние на вид кривой распределения

Для описания функции распределения пользуются специальными мерами, которые позволяют охарактеризовать положение, форме и другие особенности. Выделяют следующие меры.

Центр распределения характеризуется *средним значением* μ, *медианой Me* и *модой Mo*.

Среднее значение равно математическому ожиданию случайной величины. Данная величина также называется ***первым начальным моментом***. Первый начальный момент указывает на центр тяжести в геометрии распределения.

$$M_x = R_1 = \frac{1}{N}\sum_{i=1}^{N} x_i = \int_{-\infty}^{+\infty} x\, f(x)dx. \tag{2.18}$$

Медиана делит площадь, ограниченную функцией плотности вероятности, на две равные части, то есть соответствует условию $P(X \le Me) = F(Me) = 0{,}5$.

Мода является наиболее вероятным значением случайной величины, то есть соответствует значению $\tilde{x}$, для которого $f(\tilde{x}) = \max$.

Рассеяние случайных величин вокруг центра группирования оценивается *дисперсией*, *стандартным отклонением*, *коэффициентом вариации* и *размахом*.

Дисперсия – это среднее арифметическое (т.е математическое ожидание) квадрата отклонения случайной величины от их среднего арифметического значения. Дисперсию также называют вторым моментом случайной величины.

$$D_x = R_2 = \frac{1}{N}\sum_{i=1}^{N}(x_i - M_x)^2 = \int_{-\infty}^{+\infty}(x - M_x)^2 f(x)dx. \tag{2.19}$$

Среднее квадратическое отклонение, СКО – корень из дисперсии

$$\sigma = \sqrt{D_x}. \tag{2.20}$$

Кроме этого используют понятие стандартного отклонения:

$$\sigma_{cm} = \sigma\begin{cases} 1/\sqrt{N}, & \textit{для выборки} \\ 1/\sqrt{N-1}, & \textit{для генеральной совокупности} \end{cases}. \tag{2.21}$$

Коэффициент вариации – отношение стандартного отклонения к математическому ожиданию случайной величины.

Размах является разностью между большим и меньшим элементом выборки, то есть он равен $w = x_{\max} - x_{\min}$.

Скошенность распределения, когда один хвост кривой распределения крутой, а другой - пологий, характеризует *коэффициент асимметрии*, a_3.

Коэффициент асимметрии равен третьему моменту, деленному на куб стандартного отклонения.

$$a_3 = \frac{R_3}{\sigma_{cm}^3} = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - M_x)^3}{\sigma_{cm}^3} = \frac{1}{\sigma_{cm}^3}\int_{-\infty}^{\infty}(x - M_x)^3 f(x)dx . \qquad (2.22)$$

Пример влияния коэффициента асимметрии на кривую распределения показан на рисунке 2.4. По симметричности распределения делятся на симметричные (a_3=0, синяя сплошная), с положительной асимметрией (a_3<0, черный пунктир) и с отрицательной (a_3>0, красная сплошная).

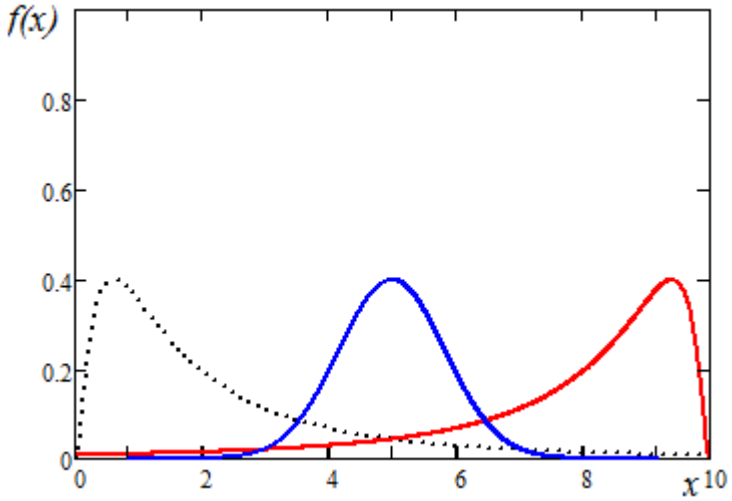


Рисунок 2.4 – Асимметрия распределений

Протяженность распределения описывается *коэффициентом эксцесса (куртозиса)* a_4. Относительное значение эксцесса равно третьему моменту, деленному на стандартное отклонение в четвертой степени. Для нормального распределения данная величина равна 3, поэтому данное значение часто «смещают» в ноль, говоря о коэффициенте эксцесса.

$$a_4 = \frac{R_4}{\sigma_{cm}^4} - 3 = \frac{\frac{1}{N}\sum_{i=1}^{N}(x_i - M_x)^4}{\sigma_{cm}^4} - 3 = \frac{1}{\sigma_{cm}^4}\int_{-\infty}^{\infty}(x - M_x)^4 f(x)dx - 3 . \qquad (2.23)$$

На рисунке 2.5 приведен пример распределений с различным коэффициентом эксцесса. Если коэффициент эксцесса a_4<0, говорят о менее протяженных, чем нормальное, распределениях (распределении с «тяжелыми» хвостами). При a_3>0 распределение является более протяженным, чем нормальное (распределение с «легкими» хвостами).

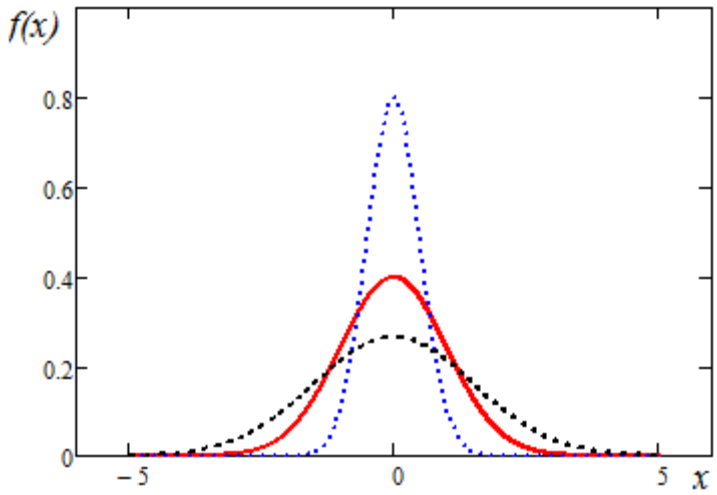


Рисунок 2.5 – Эксцесс распределения

Кроме того, одной из характеристик закона распределения является ***квантиль*** - значение, которое заданная случайная величина не превышает с фиксированной вероятностью. Т.е. квантиль можно рассматривать как обратную величину функции *F(x)*.

3 Основные законы распределения, применяемые при обработке данных научного эксперимента

В зависимости от типа случайной величины (дискретная или непрерывная) распределения также делятся на дискретные и непрерывные. Кроме того, как дискретных, так и непрерывных распределений существует довольно много, и хотя практически все из них имеют те или иные приложения в области радиотехники, рассматривать их все представляется нецелесообразным.

В данной лекции кратко рассмотрим лишь некоторые наиболее простые и распространенные из них.

3.1 Нормальное распределение

Нормальное распределение является одним из основных законов распределения в прикладной математической статистике. Во многом это проистекает из центральной предельной теоремы, которая утверждает, что распределение суммы независимых случайных величин с любым исходным распределением будет нормальным, если число слагаемых достаточно велико, а вклад каждого в сумму – мал.

Центральная предельная теорема соответствует многим реальным физическим процессам, порождающим результаты обрабатываемых наблюдений. Кроме того, при возрастании объема выборки большинство распределений стремится к нормальному, поэтому нормальное распределение может быть использовано для аппроксимации таких распределений.

В теории распространения радиоволн нормальное распределение используется для описания мерцания - флуктуаций параметра относительно его среднего значения. Кроме того, оно используется в неявном виде для описания параметра, представленного в логарифмическом масштабе (заменяя явный вид логнормального распределения).

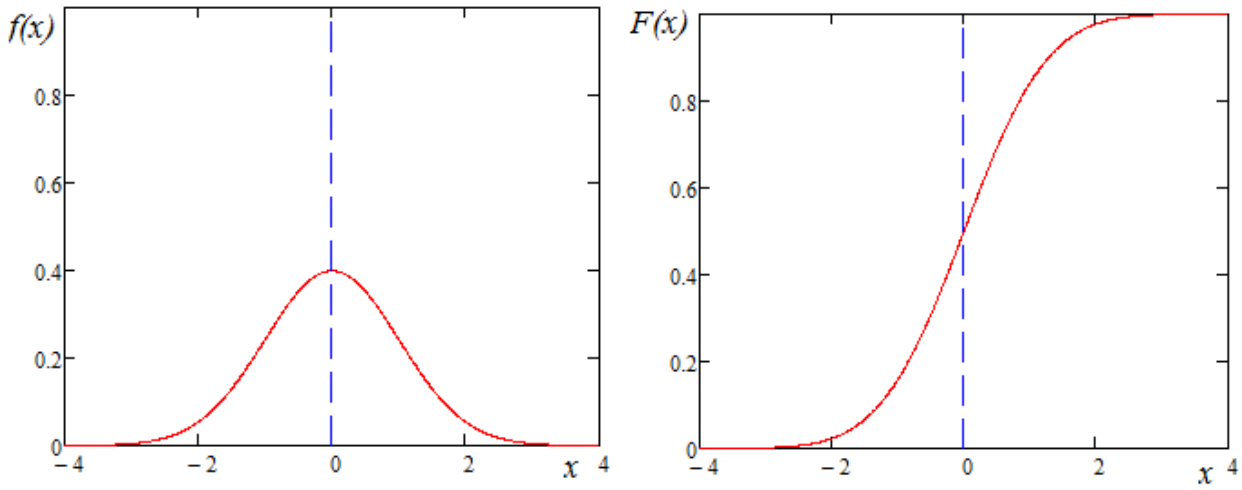
В теории надежности нормальное распределение обычно используется для описания износовых отказов, интенсивность которых со временем возрастает.

Свойства нормального закона распределения:

Обозначение	$N(\mu, \sigma)$
Параметры	μ, σ
Плотность	$f(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$
Функция распределения	$F(x;\mu,\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right)dt$
Мат. ожидание	$M(x)=\mu$
Дисперсия	$D(x)=\sigma^2$
Коэффициент вариации	$\nu=\frac{\sigma}{\mu}$
Коэффициент асимметрии	$a_3=0$

Коэффициент эксцесса	$a_4=0$
Мода	$Mo=\mu$
Медиана	$Me=\mu$

Графики $N(0,1)$:



Плотность вероятности Функция распределения

Штриховым пунктиром – линия моды и медианы

Квантиль u_p нормально распределенной случайной величины $N(\mu, \sigma)$ связан с квантилью случайной величины, имеющей стандартное нормальное распределение $N(0,1)$, u_p^c, соотношением $u_p = \mu + u_p^c$. Значения квантиля u_p^c задается таблично.

Нормальное распределение $N(\mu, \sigma)$ симметрично относительно точки $x=\mu$ и имеет два параметра, совпадающих со средним значением и стандартным отклонением. Параметры μ и σ имеют смысл коэффициента сдвига и масштаба соответственно.

Значения интегральной функции закона стандартного нормального распределения $N(0,1)$ приводятся в справочниках виде таблиц. Также приводятся таблицы значений функции (интеграла) Лапласа

$$\Phi(z) = 2F(z) - 1, \quad z = \frac{x-\mu}{\sigma}. \tag{3.1}$$

Для стандартного нормального распределения

$$F(-z) = 1 - F(z). \tag{3.2}$$

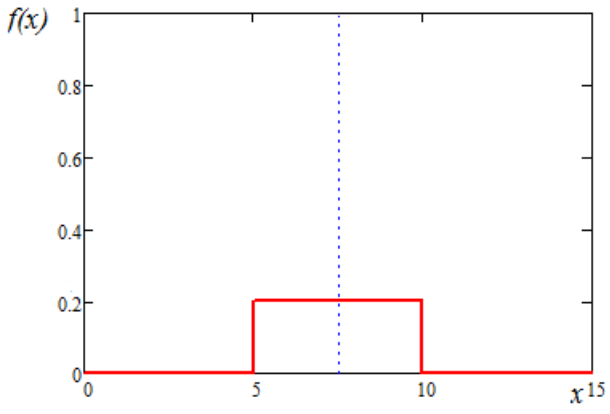
3.2 Равномерное распределение

Равномерному распределению подчиняются случайные величины, имеющие одинаковую вероятность появления (например, погрешность измерений с округлением).

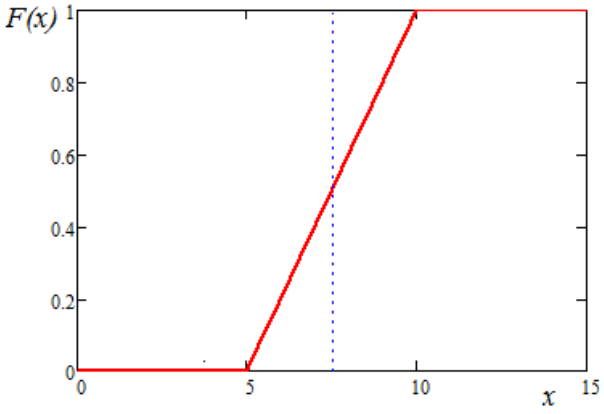
Свойства равномерного распределения:

Обозначение	$U(a, b)$
Параметры	a, b
Плотность	$f(x;a,b)=\begin{cases}\frac{1}{b-a}, & a<x<b\\ 0, & x<a; x>b\end{cases}$
Функция распределения	$F(x;a,b)=\begin{cases}0, & x\le a\\ \frac{x-a}{b-a}, & a<x<b\\ 1, & x\ge b\end{cases}$
Мат. ожидание	$M(x)=\frac{b+a}{2}$
Дисперсия	$D(x)=\frac{(b-a)^2}{12}$
Коэффициент вариации	$v=\frac{1}{\sqrt{3}}\frac{b-a}{b+a}$
Коэффициент асимметрии	$a_3=0$
Коэффициент эксцесса	$a_4=-1{,}2$
Мода	$Mo=\frac{b+a}{2}=M(x)$
Медиана	не определена

Графики $R(5,10)$:



Плотность вероятности



Функция распределения

Штриховым пунктиром – линия моды

Сумма n независимых равномерно распределенных случайных величин описывается нормальным распределением уже при $n\ge5$. Функция распределения любой случайной величины y-$F(y)$ сама распределена равномерно на отрезке [0,1].

3.3 Логарифмически нормальное распределение

Если случайная величина Y распределена нормально, то случайная величина $x=\ln(Y)$ подчинена логарифмически нормальному (логнормальному) закону распределения.

Значения логарифмически нормальной случайной величины формируются под воздействием очень большого числа взаимно независимых факторов, причем воздействие каждого отдельного фактора «равномерно незначительно» и равновероятно по знаку. При этом в отличие от нормального закона последовательный характер воздействия случайных факторов таков, что случайный прирост, вызываемый действием каждого следующего фактора, пропорционален уже достигнутому к этому моменту значению исследуемой величины.

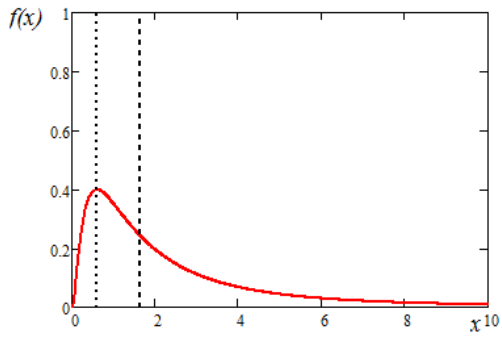
Логарифмически нормальное распределение очень часто используется в задачах распространения радиоволн, главным образом для описания параметров, связанных с мощностью, напряженностью поля, или со временем. Мощность или напряженность поля выражаются в основном только в децибелах, поэтому в этих случаях обычно о логнормальном законе говорят как о нормальном распределении уровней. В случае параметров, связанных со временем (например, длительность замираний), логнормальное распределение используется в явном виде, поскольку переменной величиной является секунда или минута, а не их логарифм.

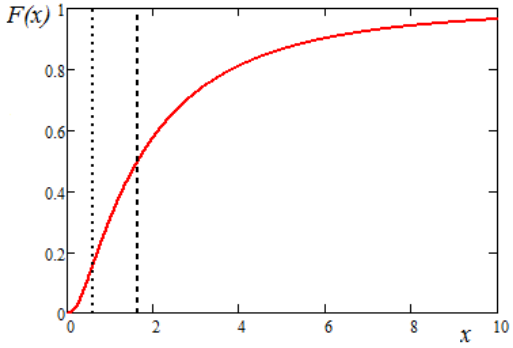
Кроме того, логнормальное распределение часто используется для описания износовых отказов. У многих невосстанавливаемых электронных приборов (некоторые типы электронных ламп, полупроводниковые приборы) наработка на отказ распределена логнормально.

Свойства логнормального распределения:

Обозначение	$LN(\mu,\sigma)$
Параметры	μ,σ
Плотность	$f(x;\mu,\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}\exp\left(-\frac{1}{2}\left(\frac{\ln x-\mu}{\sigma}\right)^2\right)$
Функция распределения	$F(x;\mu,\sigma)=\frac{1}{x\sigma\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{1}{2}\left(\frac{\ln t-\mu}{\sigma}\right)^2\right)dt$
Мат. ожидание	$M(x)=\exp\left(\mu+\frac{1}{2}\sigma^2\right)$
Дисперсия	$D(x)=\exp\left(2\mu+\sigma^2\right)\left(e^{\sigma^2}-1\right)$
Коэффициент вариации	$v=\left(e^{\sigma^2}-1\right)^{0,5}$
Коэффициент асимметрии	$a_3=\left(e^{\sigma^2}-1\right)^{0,5}\left(e^{\sigma^2}+2\right)$
Коэффициент эксцесса	$a_4=\left(e^{\sigma^2}-1\right)\left(e^{3\sigma^2}+3e^{2\sigma^2}+6\right)$
Мода	$Mo=\exp\left(\mu-\sigma^2\right)$
Медиана	$Me=e^{\mu}$

Графики $LN(0,5;1)$:



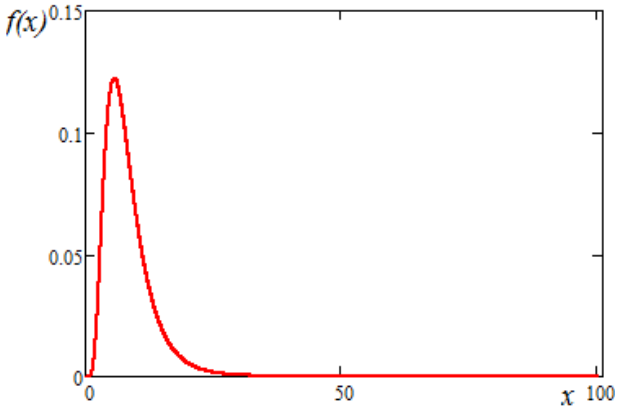


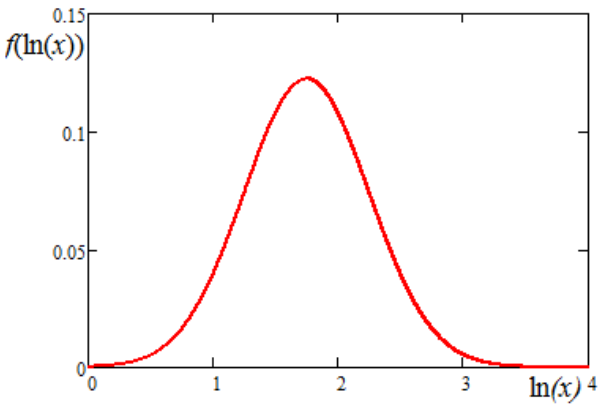
Плотность вероятности Функция распределения

Точечным пунктиром – мода, штриховым – медианы

График плотности вероятности логнормального закона распределения может быть преобразован в график плотности вероятности для нормального закона распределения, если в качестве значений случайной величины взять натуральный логарифм ее значений.

Графики плотности вероятности *LN*(2; 0,5):





Асимметрия положительна. Произведение независимых случайных величин, подчиняющихся логнормальному закону, также логнормально.

При вычислениях, связанных с логнормальным распределением, пользуются приемами для нормального распределения, заменяя при этом значение случайной величины ее логарифмом.

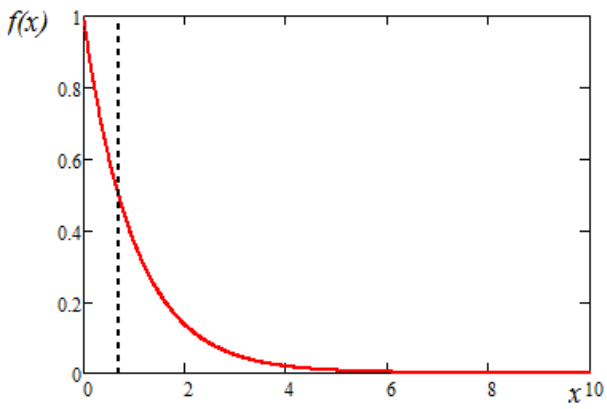
3.4 Экспоненциальное распределение

Экспоненциальное распределение - одно из наиболее часто встречающихся распределений в теории надежности. Используется для описания внезапных отказов, когда износом изделия можно пренебречь. Наработка на отказ многих невосстанавливаемых изделий и наработка между соседними отказами у восстанавливаемых изделий в случае простейшего потока подчинены этому распределению. Наработка на отказ большой многокомпонентной системы может быть описана экспоненциальным распределением при любом распределении наработки на отказ компонентов системы.

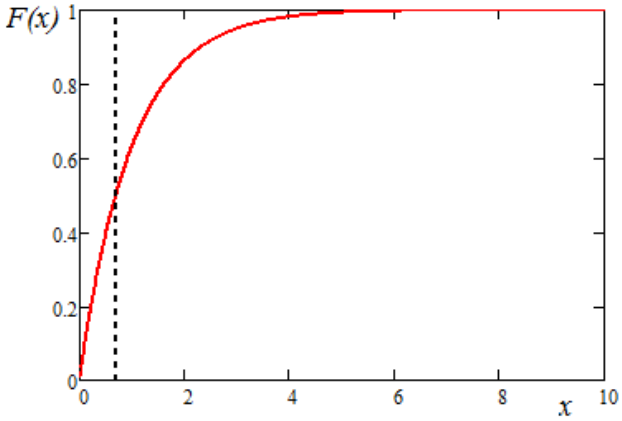
Свойства экспоненциального распределения.

Параметры	b
Плотность	$f(x;b)=\frac{1}{b}\exp\left(-\frac{x}{b}\right),\ x\geq 0$
Функция распределения	$F(x;b)=1-\exp\left(-\frac{x}{b}\right),\ x\geq 0$
Мат. ожидание	$M(x)=b$
Дисперсия	$D(x)=b^2$
Коэффициент вариации	$v=1$
Коэффициент асимметрии	$a_3=2$
Коэффициент эксцесса	$a_4=6$
Мода	$Mo=0$
Медиана	$Me=b\ln 2$

Графики экспоненциального закона распределения с параметром b = 1:



Плотность вероятности



Функция распределения

Штриховым пунктиром – медиана

Коэффициент b имеет смысл коэффициента масштаба.

Экспоненциальное распределение – частный случай распределения Вейбулла. Отличительная особенность – постоянство интенсивности отказов $\lambda=1/b=\text{const}$ – в теории надежности интерпретируется как независимость вероятности отказа от наработки, что эквивалентно отсутствию износа.

3.5 Распределение Вейбулла

Распределению Вейбулла подчиняется наработка на отказ многих невосстанавливаемых электронных приборов (электронные лампы, полупроводниковые приборы, некоторые приборы СВЧ). Распределение характеризуется разнообразием форм кривых распределения.

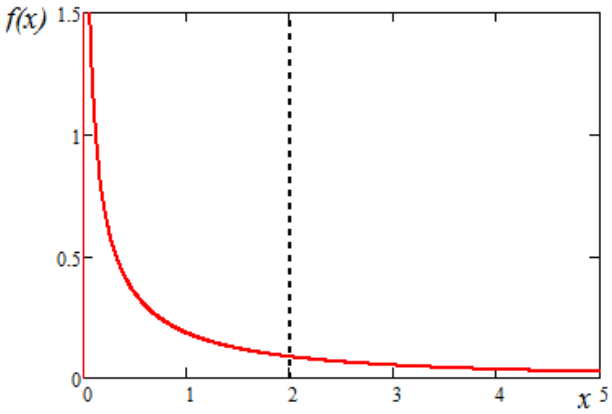
Свойства распределения Вейбулла.

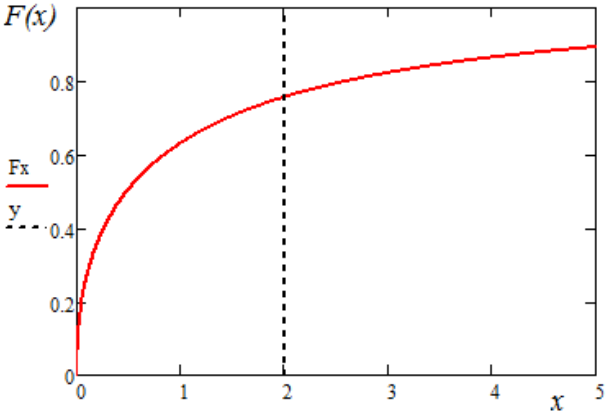
Обозначение	$W(\alpha, \beta, \mu)$
Параметры	α – коэффициент масштаба; β – коэффициент формы; μ – коэффициент сдвига

Плотность	$f(x;\alpha,\beta,\mu)=\frac{\beta}{\alpha^{\beta}}(x-\mu)^{\beta-1}\exp\left(-\left(\frac{x-\mu}{\alpha}\right)^{\beta}\right),\ x\geq\mu,\ \ \alpha,\beta>0$
Функция распределения	$F(x;\alpha,\beta,\mu)=1-\exp\left(-\left(\frac{x-\mu}{\alpha}\right)^{\beta}\right),\ x\geq\mu,\ \ \alpha,\beta>0$
Мат. ожидание	не приводим, т.к. величины имеют сложные формулы.
Дисперсия	
Коэффициент вариации	
Коэффициент асимметрии	
Коэффициент эксцесса	
Мода	$Mo=\alpha\left(1-\frac{1}{\beta}\right)^{\frac{1}{\beta}},\ \beta>1$

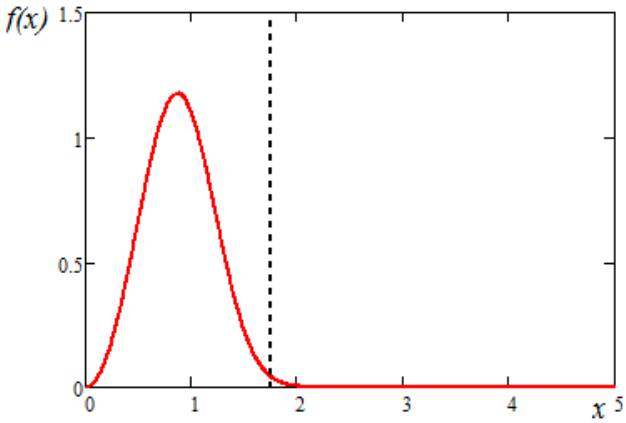
Графики закона распределения Вейбулла:

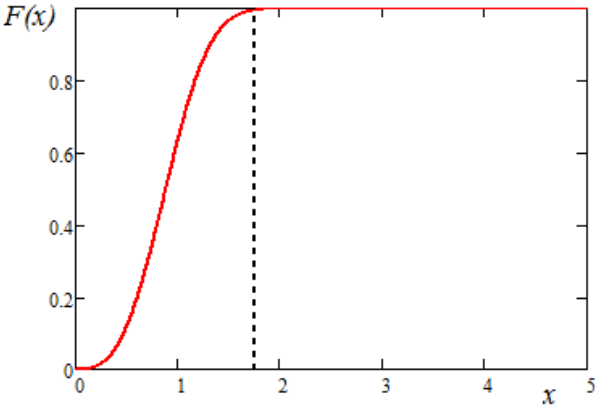
И(2, 0,5, 0)





И(2,3,0)





Плотность вероятности Функция распределения

Штриховым пунктиром – мода.

Семейство форм кривых распределения характеризуется интенсивностью отказа – функцией вида

$$\lambda(x)=\frac{f(x)}{1-F(x)},\qquad (1.7)$$

равной

$$\lambda(x)=\frac{\beta}{\alpha^{\beta}}x^{\beta-1}. \qquad (1.8)$$

Весь интервал времени жизни изделия можно разбить на три периода:

- период приработки (обкатки). Интенсивность отказа $\lambda(x)$ имеет высокие значения и явную тенденцию к убыванию (чаще всего $\lambda(x)$ монотонно убывает), что объясняется наличием в партии изделий с явными и скрытыми дефектами, которые приводят к относительно быстрому выходу их из строя. На период приработки обычно распространяется гарантийный срок;

- период нормальной эксплуатации. Он характеризуется приблизительно постоянной и сравнительно низкой интенсивностью отказов. Природа отказов в этот период носит внезапный характер (аварии, ошибки эксплуатационных работников и т.п.) и не зависит от длительности эксплуатации изделия;

- период старения и износа. Природа отказов в этот период - в необратимых физико-механических и химических изменениях материалов, приводящих к прогрессирующему ухудшению качества единицы продукции и окончательному выходу ее из строя.

Каждому периоду соответствует свой вид функции $\lambda(x)$. Значения $\beta \in (0,1)$, $\beta = 0$ и $\beta > 1$ характеризуют интенсивность отказов в периоды приработки, нормальной эксплуатации и старения соответственно.

Экспоненциальное распределение – частный случай распределения Вейбулла, соответствующее значению $\beta = 1$.

4 Определение закона распределения по выборке. Графические способы

4.1 Определение закона распределения по выборочным данным

Полученные в ходе многократного эксперимента (эмпирические) данные представляют собой с точки зрения математической статистики выборку значений случайной величины. Сама случайная величина распределена по некоторому закону, который исследователю необходимо найти, чтобы, например, получить информацию о физическом смысле рассматриваемого явления или процесса. Задача нахождения закона распределения случайной величины по выборке ее значений является одной из центральных задач математической статистики.

Существует два класса методов определения закона распределения случайной величины по выборочным данным:

- графические методы, которые представляют собой графики, отображающие значения некоторых характеристик выборки. Сюда относятся гистограмма, полигон частот, ящик с усами, вероятностные графики. По виду графиков можно сделать приблизительный вывод о виде закона распределения;

- математические методы, куда относятся критерии согласия, о которых мы поговорим в следующей лекции.

Алгоритм построения гистограммы был разобран на лекции 2. Приведем его еще раз для выборки $x_1, x_2, \ldots, x_N$ объема N:

1. построение вариационного ряда

$$x_1 \le x_2 \le \ldots \le x_N; \qquad (4.1)$$

2. разбиение интервала $[x_1, x_N]$ на n «карманов» некоторым способом;
3. нахождение границ «карманов»;
4. нахождение числа точек T_i, попавших в каждый из карманов;
5. нормировка высот столбцов гистограммы;
6. построение «ступенчатого» графика: по оси абсцисс откладываются «карманы», по оси ординат – нормированные значения.

Если вместо ступенчатого графика строится ломанная линия, соединяющая высоты «столбиков», то получается полигон относительных частот.

Вид гистограммы сильно зависит от размера и числа «карманов», способа разбиения. Кроме того, по данному графику часто сложно судить не только о конкретном виде закона распределения, но даже о его характеристических свойствах, например, об асимметрии. Например, на рисунке 4.1 приведена гистограмма, построенная по выборке из генеральной совокупности, распределенной по нормальному закону распределения, однако на графике заметна некоторая асимметрия данных.

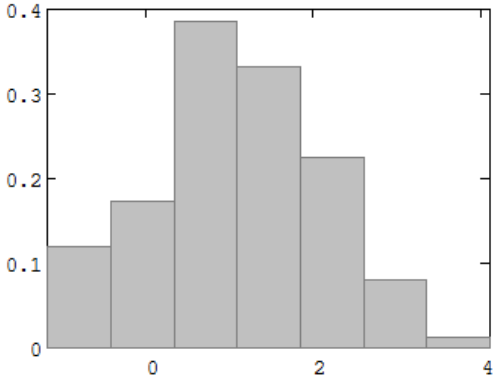


Рисунок 4.1 – Гистограмма выборки случайной величины, распределенной по нормальному закону

Больше информации о возможном виде распределения случайной величины можно получить при построении так называемых вероятностных графиков.

4.2 Виды отклонений от теоретического закона распределения на вероятностных графиках

Вероятностные графики позволяют проверить предположение о виде распределения - наглядно сравнить эмпирическое распределение с теоретическим. Если теоретическое распределение было выбрано неправильно, и в результате построения графиков выяснилось, что распределения не совпадают, то с помощью интерпретации построенного графика можно сузить область поиска следующего теоретического приближения.

При отличающихся эмпирическом и теоретическом законах распределения на вероятностных графиках возможны следующие ситуации:

- различна асимметрия распределений;
- различны толщины хвостов распределений;
- бимодальность эмпирического распределения.

Напомним, что асимметрия определяет скошенность распределения, когда один хвост графика плотности вероятности крутой, а другой - пологий. По скошенности распределения делятся на симметричные (рисунок 4.2, синяя сплошная), с положительной асимметрией, когда правый хвост длиннее левого, (рисунок 4.2, черный пунктир) и с отрицательной асимметрией, когда левый хвост длиннее правого (рисунок 4.2, красная сплошная).

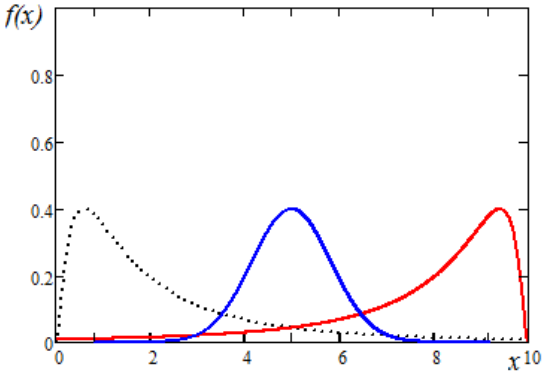
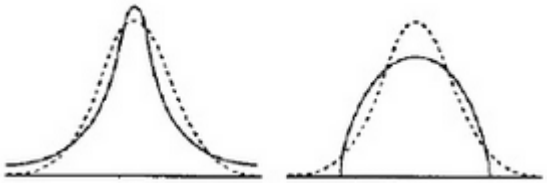


Рисунок 4.2 – Асимметрия распределений

Примером симметричного распределения может служить нормальный закон распределения.

Хвосты распределений бывают «легкими» и «тяжелыми». Говорят, что хвосты распределения «легкие», если они содержат всего лишь несколько значений. На графике плотности вероятности эти хвосты тонкие. Наоборот, «тяжелые» хвосты содержат довольно много значений и на графике выглядят толстыми. Примеры распределений с тяжелыми и легкими хвостами приведены на рисунке 4.3.



а) б)

Рисунок 4.3 – Хвосты распределений: а) распределение с «легкими» хвостами; б) распределение с «тяжелыми» хвостами

Модой распределения называется наиболее вероятное значение случайной величины. Мода нормального распределения совпадает с математическим ожиданием.

При бимодальности данные содержат две моды, т.е. два наиболее вероятных значения случайной величины. Это может трактоваться двумя способами:

- выборка не является однородной и наблюдения порождены двумя или более "наложенными" распределениями;
- выбранные инструменты не подходят для измерения.

Вид функции плотности вероятности бимодального распределения приведен на рисунке 4.4.

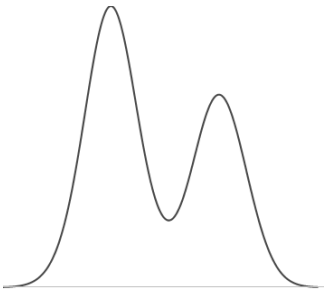


Рисунок 4.4 – Бимодальность на графике плотности вероятности

4.3 P-P график

Вероятностно-вероятностный график (процент-процентный график, *P-P* график, *probability-probability plot*) позволяет сравнить вероятности, полученные опытным путем и с помощью теоретической функцией распределения. При построении этого графика по оси ординат, откладываются теоретические значения вероятностей, по оси абсцисс – эмпирические.

P-P график строится строго на интервалах [0, 1] по обеим осям.

Рассмотрим ***алгоритм построения P-P графика*** по данным выборки $x_1, x_2, \ldots, x_N$ объема *N* в предположении, что теоретический закон распределения - нормальный:

1. построение вариационного ряда по формуле (4.1)
2. расчет точек графика p_i, i=1..*N*. Общая формула:

$$p_i = \frac{i-a}{N+1-2a}, \qquad (4.2)$$

где *a* – некоторое число от 0 до 0,5.

В зависимости от величины *a* выделяют следующие методы расчета:

- метод Блома (*Blom*)

$$p_i = \frac{i - 0{,}375}{N + 0{,}25}, \quad (4.3)$$

- метод *Тьюки* (*Tukey*)

$$p_i = \frac{i - 1/3}{N + 1/3}, \quad (4.4)$$

- метод Ван дер Вардена (*Van der Waerden*)

$$p_i = \frac{i}{N + 1}, \quad (4.5)$$

- метод рангового преобразования

$$p_i = \frac{i - 0{,}5}{N}. \quad (4.6)$$

и другие методы;

3. приведение значений вариационного ряда к стандартизированному виду построением вариационного ряда Z_i по формуле;

$$Z_i = \frac{X_i - \mu}{\sigma}, \quad (4.7)$$

где μ – среднее арифметическое выборки;

σ – стандартное отклонение выборки;

4. расчет Ф(Z_i) - значений функции Лапласа по значениям Z_i;

5. нанесение на график точек с координатами (p_i, Ф(Z_i)).

4.4 Q-Q график

Квантиль-квантильный график (*Q-Q* график, *quantile-quantile plot*) позволяет сравнить значения квантилей эмпирической функции распределения, которыми являются данные эксперимента, и теоретической функциии распределения. Напомним, что под квантилью понимается значение, которое заданная случайная величина не превышает с фиксированной вероятностью. Т.е. квантиль можно рассматривать как обратную величину функции *F(x)*.

По оси ординат *Q-Q* графика откладываются экспериментальные значения, по абсцисс квантили теоретического распределения.

Рассмотрим ***алгоритм построения Q-Q графика*** по данным выборки $x_1, x_2, \ldots, x_N$ объема *N* в предположении, что теоретический закон распределения - нормальный:

1. построение вариационного ряда по формуле (4.1);

2. расчет точек графика p_i, i=1..*N* по любой из формул (4.3)-(4.6);

3. расчет значений квантилей теоретического закона распределения Ф в точках разбиения по формуле

$$Q_i = \Phi^{-1}(p_i); \quad (4.8)$$

4. нанесение на график точки с координатами (Q_i, X_i).

4.5 Интерпретация вероятностных графиков

1) Если *P-P* график образует прямую, проведенную через точку (0,0) под углом 45°, говорят, что эмпирическое и теоретическое распределения совпадают.

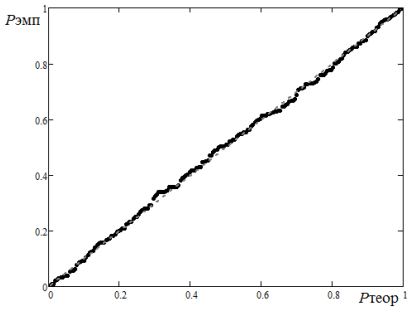


Рисунок 4.5 – P-P график при совпадении законов распределения

2) Q-Q график образует прямую вида $y=ax+b$, считается, что экспериментальные данные распределены по теоретическому закону распределения со средним, равным коэффициенту *b*, и стандартным отклонением, равным коэффициенту *a*. Например, точки графика практически совпадают с прямой с коэффициентами a=0,023, b=9,261 (рисунок 4.6). Следовательно, экспериментальные данные распределены по нормальному закону N(9,261, 0,023).

Напомним, что коэффициент *a* представляет собой тангенс угла наклона прямой, а коэффициент *b* – сдвиг графика по оси ординат.

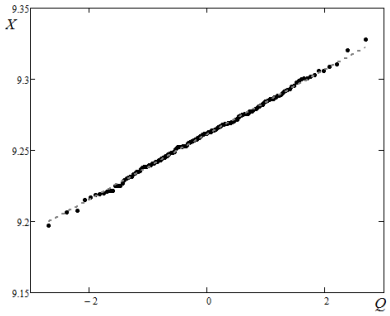
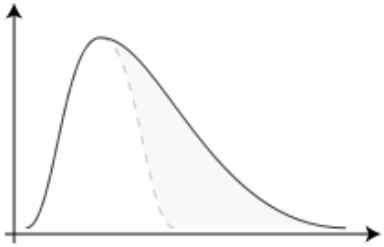


Рисунок 4.6 – Квантиль-квантильный график данных, распределенных по закону N(9,261, 0,023)

Дальнейшая интерпретация графиков связана с отклонениями точек от прямой и для *P-P* и *Q-Q* полностью совпадает.

3) Если полученный график похож на выгнутую вниз дугу (рисунок 4.7, а), то говорят, что эмпирические данные имеют положительную асимметрию (рисунок 4.7, б).

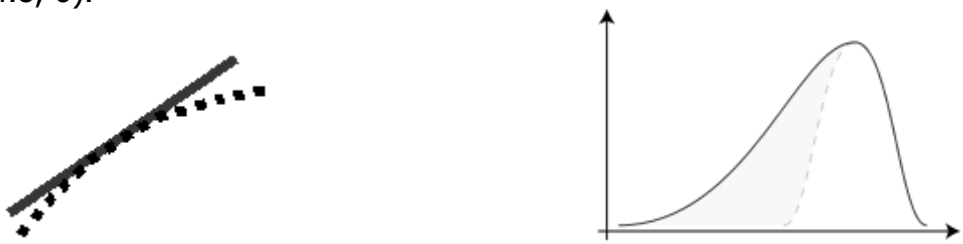
а)



б)

Рисунок 4.7 – Положительная асимметрия эмпирических данных

4) Если полученный график похож на выгнутую вверх дугу (рисунок 4.8, а), то говорят, что эмпирические данные имеют отрицательную асимметрию (рисунок 4.8, б).



а) б)

Рисунок 4.8 –Отрицательная асимметрия эмпирических данных

5) Если полученный график повторяет форму, приведенную на рисунке 4.9, а, то говорят о тяжелых хвостах эмпирического распределения (рисунок 4.9, б).

а) б)

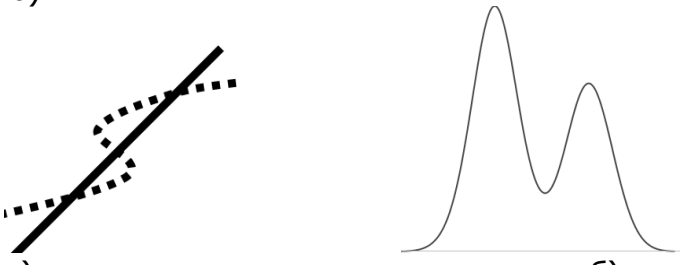
Рисунок 4.9– Тяжелые хвосты эмпирического распределения

6) Если полученный график повторяет форму, приведенную на рисунке 4.10, а, то говорят о легких хвостах эмпирического распределения (рисунок 4.10, б).

а) б)

Рисунок 4.10 –Легкие хвосты эмпирического распределения

7) Если полученный график повторяет форму, приведенную на рисунке 4.11, а, то говорят о бимодальности эмпирического распределения (рисунок 4.11, б).



а) б)

Рисунок 4.11 – Бимодальное распределение эмпирических данных

У *P-P* графиков лучше разрешение (то есть нагляднее отклонения) по центру графика, а у *Q-Q* – по краям. Поскольку основные виды отклонений (рисун-

ки 4.7-4.10) сконцентрированы по краям, то чаще используется *Q-Q* график. Вместе с тем различия, характеризующие бимодальность, сконцентрированы в центре графика и следовательно, будут лучше всего видны на *P-P* графике.

5 Определение закона распределения по выборке. Критерии согласия

5.1 Общий алгоритм проверки статистических гипотез

Пусть в ходе эксперимента исследуется некоторая случайная величина *X*. В результате исследования была получена выборка ее значений объема N: $x_1, x_2, \ldots, x_N$. Любое предположение о распределении ее вероятностей называется статистической гипотезой. То есть ***статистическая гипотеза*** *– это определенное предположение о распределении вероятностей, лежащем в основе наблюдаемой выборки данных*.

Процесс принятия решения о том, противоречит ли рассматриваемая статистическая гипотеза наблюдаемой выборке или нет, называется проверкой статистической гипотезы. Данная проверка осуществляется на основе статистических критериев.

Исходную гипотезу называют нулевой гипотезой и обозначают H_0. Во время проверки одновременно рассматривается противоречащая ей гипотеза, которая называется конкурирующей или альтернативной и обозначается H_1.

При проверке статистических гипотез применяется следующий ***общий алгоритм***:

1. Формулируются нулевая H_0 и альтернативная H_1 гипотезы о распределении вероятностей на множестве значений случайной величины *X* (генеральной совокупности). Иногда альтернатива формулируется не в явном виде, тогда предполагается, что H_1 означает «не H_0».

2. Задается некоторая статистика (функция выборки): $T = T(x_1, \ldots x_N)$, которая в условиях справедливости гипотезы H_0 подчиняется некоторому известному закону распределения. Часто вопрос о том, какую статистику выбрать, не имеет однозначного ответа.

3. Задается уровень значимости α – допустимая для данной задачи вероятность ошибки первого рода. На практике часто полагают α=0,05.

4. На множестве допустимых значений статистики выделяется критическая область *C* наименее вероятных значений статистики. Вероятность того, что значение статистики попадет в эту область при справедливости нулевой гипотезы, равна α, т.е $P(T \in C \mid H_0) = \alpha$. Задача вычисления границ критического множества в большинстве практических случаев имеет готовое решение.

5. Статистический критерий заключается в проверке условия:

- если $T(x_1, \ldots x_N) \in C$, то делается вывод, что данные противоречат нулевой гипотезе при уровне значимости α. H_0 отвергается, H_1 принимается;
- если $T(x_1, \ldots x_N) \notin C$, то делается вывод, что данные не противоречат нулевой гипотезе при уровне значимости α. H_0 принимается, H_1 отвергается.

Таким образом, статистический критерий определяется статистикой *T* и критической областью *C*, которая зависит от уровня значимости α.

Если данные не противоречат нулевой гипотезе, то это еще не значит, что гипотеза верна:

- для нахождения различий не хватает объема выборки;
- выбранная статистика отражает не всю информацию нулевой гипотезы и это искажает процесс принятия решения.

Выделяют три вида критических областей:

1. двусторонняя критическая область определяется двумя интервалами $(-\infty, x_{\alpha/2}) \cup (x_{1-\alpha/2}, +\infty)$, где точки $x_{\alpha/2}$ и $x_{1-\alpha/2}$ находят из условий

$$P(T < x_{\alpha/2}) = \frac{\alpha}{2}; \quad (5.1)$$

$$P(T < x_{1-\alpha/2}) = 1 - \frac{\alpha}{2}; \quad (5.2)$$

2. левосторонняя критическая область определяется интервалом $(-\infty, x_{\alpha})$, где x_{α} находят из условия

$$P(T < x_{\alpha}) = \alpha; \quad (5.3)$$

3. правосторонняя критическая область: $(x_{1-\alpha}, +\infty)$, где $x_{1-\alpha}$ находят из условия

$$P(T < x_{1-\alpha}) = 1 - \alpha. \quad (5.4)$$

Как было отмечено выше, при проверке статистической гипотезы возможны следующие ситуации:

		Верная гипотеза	
		H_0	H_1
Результат применения критерия	H_0	H_0 верно принята	H_0 неверно принята (ошибка II рода)
	H_1	H_0 неверно отвернута (ошибка I рода)	H_0 верно отвернута

Ошибки первого и второго родов являются взаимно симметричными, то есть, если поменять местами гипотезы H_0 и H_1, то ошибка первого рода становится ошибкой второго и наоборот.

Ошибку первого рода, когда верная гипотеза отвергается, часто называют ложной тревогой, ложным (ложноположительным) срабатыванием. Уровень значимости α задает вероятность совершить данную ошибку.

Ошибка второго рода называется пропуском события или ложноотрицательным срабатыванием. Вероятность совершения данного рода ошибки определяется величиной β. С данной величиной связано понятие мощности критерия 1-β, которая тем выше, чем меньше вероятность совершить ошибку второго рода.

При применении статистических критериев приходится идти на компромисс между приемлемым уровнем ошибок первого и второго уровней.

Рассмотрим один из классов статистических критериев: критерии согласия. Данные критерии проверяют, согласуется ли заданная выборка с некоторым теоретическим распределением.

Статистические гипотезы различают по виду предположений, которые в них содержатся:

• ***простые***, *когда выдвигается предположение не только о виде распределения, но и о значениях всех его параметров*. Например, в качестве теоретического распределения выбран закон $N(0,1)$;

• ***сложные***, *когда выдвигается предположение о виде закона распределения (например, нормальный, экспоненциальный), при этом один или несколько параметров в предположении неизвестны*. Например, в качестве теоретического закона предполагается нормальный закон с неизвестными значениями обоих

параметров $N(\mu,\sigma)$, либо с неизвестным значением μ, либо с неизвестным значением σ.

Среди критериев согласия для проверки гипотезы о виде эмпирического распределения наиболее часто используются критерий Колмогорова и критерий Пирсона (критерий «хи-квадрат», χ^2). Если про выборку заранее известно, что она подчиняется нормальному закону, становится возможно применять более мощные критерии нормальности – частный случай критериев согласия.

5.2 Критерий согласия Колмогорова

Данный критерий Колмогорова в классическом виде используется для проверки простой гипотезы, поскольку в этом случае он является «свободным от распределения» (то есть статистика не зависит от вида наблюдаемого закона распределения и его параметров). В случае проверки сложной гипотезы применение критерия Колмогорова требует соблюдения некоторых дополнительных условий.

В критерии Колмогорова сравнение проводится между теоретической и эмпирической функциями распределения, как показано на рисунке 5.1. Критерий позволяет найти точку, в которой расхождение между этими двумя функциями является наибольшим, и оценить достоверность этого расхождения.

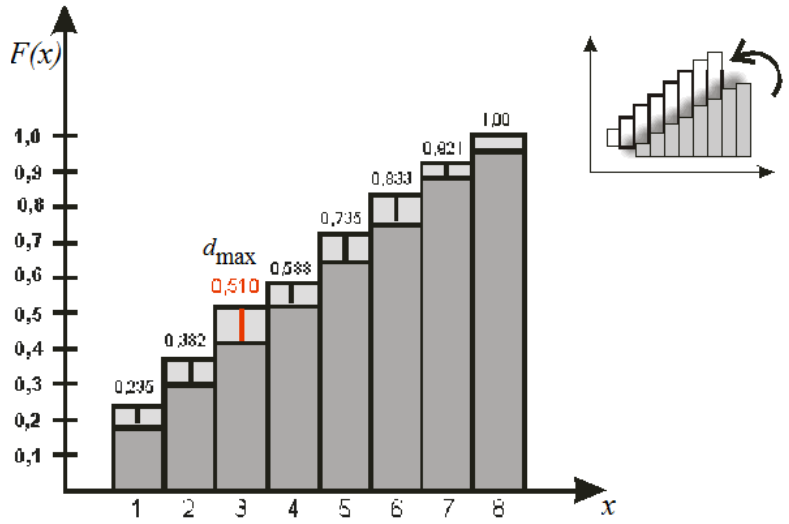


Рисунок 5.1 – Графическое представление критерия согласия Колмогорова

К достоинствам критерия Колмогорова по сравнению с другими критериями согласия можно отнести его простоту. Из недостатков можно выделить следующее:

- в чистом виде неприменим для доказательства сложных гипотез;
- обладает низкой чувствительностью. Во-первых, при небольших объемах экспериментальных данных (N<50) имеет небольшую достоверность и может подтвердить совпадение явно несовпадающих визуально распределений. Во-вторых, даже при больших объемах экспериментальных данных может подтвердить совпадение несовпадающих распределений.

Рассмотрим ***алгоритм применения критерия Колмогорова при проверке простой гипотезы***. Пусть имеется выборка из N>20 экспериментальных значений величины X.

1 H_0: X распределена по некоторому закону $F(x, \theta)$, θ – известные значения параметров.

2 Строится вариационный ряд (сортировка данных)

$$x_1 \le x_2 \le \ldots \le x_N . \quad (5.5)$$

3 Строится эмпирическая функция распределения

$$F_j^{emp} = \frac{1}{N}\sum_{i=1}^{N} I_{x_i < x_j}, \quad j = 1 \ldots N, \quad (5.6)$$

где I – частота попадания случайной величины x_i в интервал $(-\infty; x_j]$:

$$I_{x_i \le x_j} = \begin{cases} 1, & x_i \le x_j; \\ 0, & x_i > x_j. \end{cases} \quad (5.7)$$

4 В точках x_i строится теоретическая функция распределения F^{teor}

$$F_j^{teor} = F(x_j, \theta). \quad (5.8)$$

5 Вычисляется наибольшая разница между значениями F^{emp} и F^{teor} в точках x_i:

$$D = \max_j \left| F_j^{emp} - F_j^{teor} \right| \quad (5.9)$$

и статистика критерия Колмогорова вида

$$K_{набл} = \sqrt{N} D . \quad (5.10)$$

6 Задается доверительная вероятность P (обычно 0,95) или уровень значимости α=1-P. По α находится критическое значение критерия Колмогорова $К_{кр}$ (табличное или по приближенной формуле)

$$K_{кр} \approx \sqrt{-\frac{1}{2}\ln\frac{\alpha}{2}} . \quad (5.11)$$

7 Если

$$K_{набл} < K_{кр}, \quad (5.12)$$

то гипотеза H_0 принимается.

При проверке сложной гипотезы на закон распределения статистики Колмогорова влияют следующие факторы:

- сближение выборочное и теоретическое распределений из-за оценки неизвестных параметров по выборке, таким образом, значение статистики - меньше, чем в случае проверки простой гипотезы;
- вид наблюдаемого закона распределения $F(x, \theta)$;
- тип оцениваемого параметра и число оцениваемых параметров;
- в некоторых ситуациях конкретное значение параметра (например, для гамма-распределения);
- используемый метод оценивания параметров;
- объем выборки при малых объемах. Однако уже при $N \ge 15 \ldots 20$ зависимостью от N можно пренебречь.

Пусть имеется выборка из $N>20$ экспериментальных значений величины X.

1 H_0: X распределена по некоторому закону $F(x, \theta_1, \theta_2)$, θ_1 – неизвестные значения параметров; θ_2 – известные значения параметров.

2 Составление вариационного ряда (5.5)

3 Построение эмпирической функции распределения (5.6)-(5.7).

4 Оценка неизвестных параметров. Например, методом максимального правдоподобия получаются следующие оценки параметров

- для нормального распределения:

$$\mu = \frac{1}{N}\sum_{i=1}^{N} X_i \,, \tag{5.13}$$

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(X_i - \mu)^2} \,; \tag{5.14}$$

- для логнормального распределения

$$\mu = \frac{1}{N}\sum_{i=1}^{N} \ln X_i, \tag{5.15}$$

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(\ln X_i - \mu)^2}; \tag{5.16}$$

- для экспоненциального распределения

$$b = \frac{1}{N}\sum_{i=1}^{N} X_i. \tag{5.17}$$

5 В точках x_i строится теоретическая функция распределения F^{teor} (5.8).

6 Рассчитывается статистика критерия Колмогорова (5.9)-(5.10).

7 Задается уровень доверительной вероятности Р. По таблицам скорректированных критических значений критерия для конкретного вида распределения и оцениваемых параметров находится $К_{кр}$.

8 Если выполняется неравенство (5.12), гипотеза H_0 принимается.

5.3 Критерий согласия Пирсона

Критерий согласия Пирсона (критерий «хи-квадрат», χ^2) отвечает на вопрос о том, с одинаковой ли частотой встречаются разные значения некоторой величины в рассматриваемых распределениях. Таким образом, с его помощью проводится сравнение между теоретической и эмпирической плотностями распределения (рисунок 5.2).

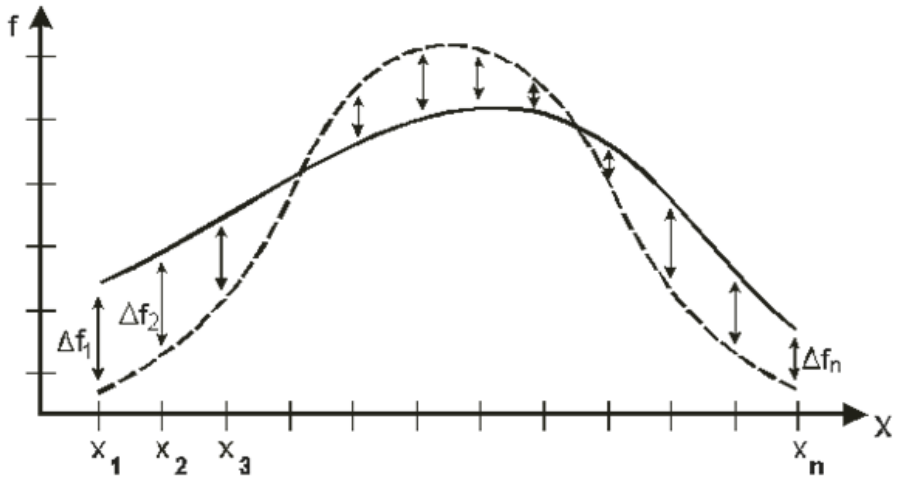


Рисунок 5.2 – Графическое представление критерия согласия Пирсона

Критерий согласия Пирсона позволяет сопоставлять распределения переменных, представленных в любой шкале, даже по шкале наименований. Данным критерием можно пользоваться даже в самом простом случае «есть результат – нет результата».

Недостатком критерия согласия Пирсона является то, что эмпирические распределения плотности вероятности, как правило, имеют очень большой разброс по интервалам. Это приводит к увеличению критерия χ^2 и отбрасыванию гипотезы о согласии даже таких распределений, которые находятся в явно хорошем согласии, если судить по графикам.

Рассмотрим алгоритм применения критерия Пирсона при проверке простой гипотезы. Пусть имеется выборка из $N>50$ экспериментальных значений величины X.

1 H_0: X распределена по некоторому закону $F(x, \theta)$, θ – известные значения параметров.

2 Строится вариационный ряд (5.5)

3 Проводится группировка данных по одному из рассмотренных ранее алгоритмов.

4 Строится эмпирическая функция плотности вероятности:

- рассчитывается число точек, попавших в каждый интервал

$$T_i = \sum_{j=1}^{N} t_{j,i}\ , \tag{5.18}$$

где T_i – количество экспериментальных точек в i-м интервале;

$t_{j,i}$ – признак наличия j-й точки в i-том интервале:

$$t_{j,i} = \begin{cases} 1, \text{ если } x_j \in [a_i, b_i], \\ 0, \text{ если } x_j \notin [a_i, b_i]. \end{cases}$$

- рассчитывается эмпирическая частота попадания в интервал:

$$q_i = \frac{T_i}{N},\ i = 1..n\ . \tag{5.19}$$

5 Рассчитываются теоретические частоты попадания в интервалы

$$p_i = P(a_i < x < b_i) = F(b_i) - F(a_i). \tag{5.20}$$

Для нормального распределения $N(\mu, \sigma)$

$$P(a_i < x < b_i) = \Phi\left(\frac{b_i - \mu}{\sigma}\right) - \Phi\left(\frac{a_i - \mu}{\sigma}\right), \tag{5.21}$$

где $\Phi(z)$ – определенная таблично функция Лапласа, являющаяся функцией распределения стандартного нормального закона $N(0,1)$.

6 Рассчитывается статистика критерия Пирсона:

$$\chi^2_{\text{набл}} = n \cdot \sum_{i=1}^{n} \frac{(q_i - p_i)^2}{p_i}\ . \tag{5.22}$$

7 Задается доверительная вероятность P или уровень значимости α и по таблицам распределений вида χ^2 находится критическое значение распределения Пирсона с уровнем значимости α и r степенями свободы, где

$$r = n - 1. \tag{5.23}$$

8 Если

$$\chi^2_{\text{набл}} < \chi^2_{\text{кр}}, \tag{5.24}$$

гипотеза H_0 принимается.

При проверке сложной гипотезы применяется тот же алгоритм со следующими отличиями:

- проводится оценка неизвестных параметров;
- число степеней свободы распределения Пирсона принимается равным

$$r = n - m - 1, \tag{5.25}$$

где m – число оцененных параметров.

6 Регрессионный анализ

6.1 Общие понятия регрессионного анализа

Пусть в ходе эксперимента были получены несколько наборов различных данных, представляющих собой значения некоторых параметров: векторов $X=\{x_1,x_2,\ldots,x_N\}$, $Y=\{y_1,y_2,\ldots,y_N\}$, и так далее. Например, X – значения температуры окружающей среды, Y - активного сопротивления электрической цепи. Выделим пару параметров X и Y и рассмотрим влияние одной величины на другую. Иногда это влияние может быть записано в явном виде с помощью некоторой функции и тогда говорят о жесткой функциональной связи между параметрами.

Регрессионный анализ занимается нахождением математического уравнения зависимости одних параметров от других – восстановлением функциональной зависимости по данным эксперимента. Искомое уравнение называется ***уравнением (функцией) регрессии***. На практике при регрессионном анализе чаще всего применяется метод наименьших квадратов.

Различают два основных типа переменных:

- ***факторы*** *- независимые переменные*;
- ***отклики*** *- зависимые переменные*.

В зависимости от возможности устанавливать значения, факторы делятся на контролируемые и управляемые и на контролируемые и неуправляемые.

В результате изменений факторов появляется эффект, который передается на отклики. Разделение на факторы и отклики не всегда четко выражено и часто зависит от целей исследователя. Так, можно рассматривать отклик промежуточной стадии процесса исследования как фактор на конечной стадии.

Любую функциональную зависимость отклика от факторов можно абстрактно представить в виде некоторого «ящика», который преобразует переменные, поступающие на вход (факторы), к переменной на выходе (отклику). В зависимости от входящих переменных функция ящика будет либо одномерной («один вход» - «один выход»), либо многомерной.

В зависимости от априорных знаний экспериментатора о структуре функции «ящика» и количественных значениях ее параметров, «ящики» делятся на «белый», «серый» и «черный», как показано в таблице 6.1

Таблица 6.1 – Классификация «ящиков»

	белый	серый	черный
структура	**+**	**+**	**-**
количественные значения параметров	**+**	**-**	**-**

С точки зрения эксперимента с «ящиком» функцией регрессии будет считаться функция «ящика», преобразующая его «входы» к «выходу».

Для нахождения функций регрессии наиболее часто используется так называемая полиномиальная модель – полином степени *n*

$$Y = f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_n X^n, \quad (6.1)$$

где β_i называются параметрами модели.

Уравнение (6.1) определяет множество всех кривых данной формы, поэтому для нахождения функции регрессии для известных значений $X=\{x_1,x_2,\ldots,x_N\}$ и $Y=\{y_1,y_2,\ldots,y_N\}$ необходимо найти неизвестные коэффициенты β_i.

Если используемая модель линейна относительно неизвестных коэффициентов (но необязательно линейна относительно независимых переменных *X*), то говорят о линейной регрессионной модели. В противном случае модель называется нелинейной.

Часто происходит так, что значения отклика несколько отклоняются от тех, которые были бы получены через функциональную зависимость (6.1). Это можно трактовать как погрешность снятия данных в эксперименте, и для соответствия данных полиномиальной модели в нее вводится вектор ошибок ε:

$$Y = f(X) = \beta_0 + \beta_1 X + \beta_2 X^2 + \ldots + \beta_n X^n + \varepsilon . \quad (6.2)$$

После нахождения функции регрессии (то есть нахождения неизвестных коэффициентов β_i) часто доказывают ***адекватность модели*** - *то, что никакая другая модель не даст значимого улучшения в предсказании отклика*.

6.2 Функция регрессии в одномерном линейном случае

Рассмотрим самый простой случай применения регрессионного анализа: одномерную линейную регрессионную модель. В данном случае под одномерностью понимается зависимость отклика от одного фактора.

Пусть вектор $X=\{X_i\}_{i=1..N}$ определяет значения фактора, а вектор $Y=\{Y_i\}_{i=1..N}$ - значения отклика. Пусть также фактор влияет на отклик почти по линейному закону.

Предположение о линейности модели относительно независимых переменных должно быть априорно проверено. Это может быть выполнено двумя способами:

- графически с помощью диаграммы рассеяния;
- математически через выборочный коэффициент линейной корреляции.

Диаграмма рассеяния – это отображение данных *X* и *Y* в виде *N* точек на декартовой плоскости с координатами (X_i, Y_i). Вид диаграммы рассеяния приведен на рисунке 6.1. Если точки практически образуют прямую, то модель считают линейной.

Векторы $X=\{x_1,x_2,\ldots,x_N\}$ и $Y=\{y_1,y_2,\ldots,y_N\}$ можно также рассматривать как выборки случайных величин *X* и *Y*. Выборочный (т.е. рассчитанный по данным выборок) коэффициент корреляции вычисляется по формуле

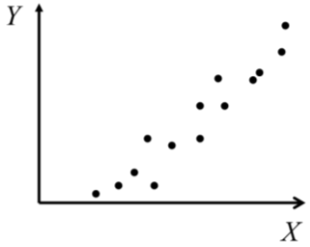


Рисунок 6.1 – Пример диаграммы рассеяния при линейной зависимости отклика от значений фактора

$$r = \frac{\sum_{i=1}^{N}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{N}(Y_i - \overline{Y})^2}}, \quad (6.3)$$

где $\overline{X}$, $\overline{Y}$ – средние значения выборок X и Y соответственно.

Считается, что между фактором X и откликом Y существует сильная линейная зависимость, если $|r| \approx 1$.

Для нахождения линейной одномерной функции регрессии на основании предварительных исследований (через диаграмму рассеяния или коэффициент линейной корреляции выборки) выдвигается гипотеза о линейной зависимости между фактором и откликом. Эта гипотеза может быть записана в виде полинома первой степени

$$Y = f(X) = aX + b + \varepsilon. \quad (6.4)$$

Ошибку ε можно трактовать как экспериментальную погрешность.

Если опустить значения ошибки ε, то (6.4) представляет собой уравнение прямой с угловым коэффициентом a и сдвигом по оси Oy, b (рисунок 6.2).

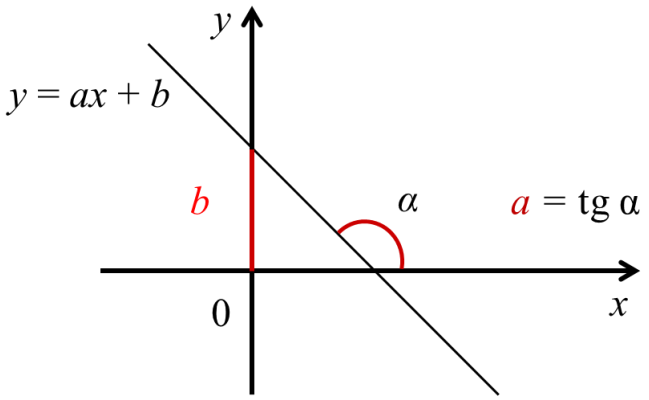


Рисунок 6.2 – Смысл коэффициентов линейной функции

Для однозначной идентификации регрессионной прямой необходимо оценить значения коэффициентов a и b по известным векторам $X=\{X_i\}$, $Y=\{Y_i\}$, $i=1\ldots N$, которые можно трактовать как данные эксперимента.

Необходимо отметить, что уравнением вида $y=ax+b$ нельзя записать прямые, параллельные оси ординат, так как $\mathrm{tg}(90°)=\infty$. Однако такие прямые не могут описывать влияние фактора на отклик, и следовательно, не рассматриваются в регрессионном анализе.

Нахождение неизвестных коэффициентов прямой осуществляется методом наименьших квадратов: строится функция суммарной ошибки величин ε_i и находятся такие значения a и b, при которых данная функция достигает минимума.

Величины ε_i можно трактовать как разность между имеющимся значением отклика Y_i и «теоретическим» значением отклика Y^T_i - точки графика функции регрессии (пока неизвестной) с координатой X_i (рисунок 6.3).

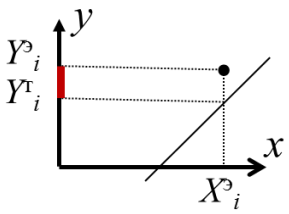


Рисунок 6.3 – К определению ошибки ε_i

Таким образом, ε_i можно записать как

$$\varepsilon_i = Y_i - Y_i^T = Y_i - (aX_i + b). \quad (6.5)$$

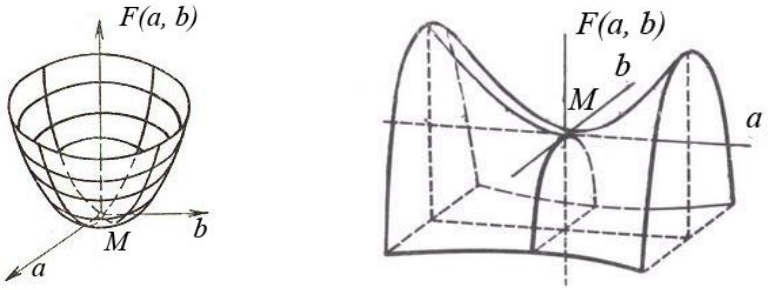
Здесь *a* и *b* – неизвестные, X_i и Y_i – известные значения. Суммарная ошибка представляет собой функцию от двух переменных *a* и *b*. Ошибки ε_i возведены в квадрат для того, чтобы избежать влияния знака погрешности (т.е. того, больше «теоретической» величины значения Y_i или меньше).

$$F(a,b) = \sum \varepsilon_i^2 . \quad (6.6)$$

При подстановке (6.5) в (6.6) последнее выражение преобразуется к виду

$$F(a,b) = \left(\sum_{i=1}^{N} X_i^2\right)a^2 + Nb^2 + \left(2\sum_{i=1}^{N} X_i\right)ab - \left(2\sum_{i=1}^{N} X_i Y_i\right)a - \left(2\sum_{i=1}^{N} Y_i\right)b + \sum_{i=1}^{N} Y_i^2 . \quad (6.7)$$

$F(a,b)$ - квадратичная функция двух переменных *a* и *b*. График подобной функции - поверхность, которая в силу квадратичности может быть либо эллиптическим параболоидом (рисунок 6.3 а), либо гиперболическим параболоидом (рисунок 6.3 б).



а) б)

Рисунок 6.3 – Параболоиды а) эллиптический, б) гиперболический

Точка минимума имеется только у эллиптического парабалоида, причем если ветви параболы, его образующей, направлены вверх. У гиперболического параболоида («седла») нет ни минимума, ни максимума, а есть лишь т.н. «седловая» точка, которая в одном сечении представляет собой локальный максимум, в другом – локальный минимум.

Для нахождения минимума функции $F(a,b)$ необходимо удовлетворить необходимым и достаточным условиям экстремума:

а) из необходимых условий экстремума

$$\begin{cases} \dfrac{\partial F}{\partial a} = 0, \\ \dfrac{\partial F}{\partial b} = 0. \end{cases} \tag{6.8}$$

находится стационарная точка *M* с координатами (*a, b*). При подстановке в (6.8) функции (6.7) и взятии производной получается система линейных алгебраических уравнений, которую можно решать любым из известных способов (результат будет одинаков). Например, методом Крамера имеем:

$$a = \frac{N\left(\sum_{i=1}^{N} X_i Y_i\right) - \left(\sum_{i=1}^{N} X_i\right)\left(\sum_{i=1}^{N} Y_i\right)}{N\left(\sum_{i=1}^{N} X_i{}^2\right) - \left(\sum_{i=1}^{N} X_i\right)^2}, \tag{6.9}$$

$$b = \frac{\left(\sum_{i=1}^{N} X_i{}^2\right)\left(\sum_{i=1}^{N} Y_i\right) - \left(\sum_{i=1}^{N} X_i Y_i\right)\left(\sum_{i=1}^{N} X_i\right)}{N\left(\sum_{i=1}^{N} X_i{}^2\right) - \left(\sum_{i=1}^{N} X_i\right)^2}; \tag{6.10}$$

б) для проверки того, что точка с координатами (*a, b*) является точкой минимума, используются достаточные условия экстремума в точке *M.* Для этого для функции двух переменных вычисляются значения

$$A = \left.\frac{\partial^2 F}{\partial a^2}\right|_M, \; B = \left.\frac{\partial^2 F}{\partial a\,\partial b}\right|_M, \; C = \left.\frac{\partial^2 F}{\partial b^2}\right|_M, \; D = AC - B^2. \tag{6.11}$$

В рассматриваемом случае

$$A = \frac{\partial^2 F}{\partial a^2} = 2\left(\sum_{i=1}^{N} X_i{}^2\right), \tag{6.12}$$

$$B = \frac{\partial^2 F}{\partial a\,\partial b} = 2\left(\sum_{i=1}^{N} X_i\right), \tag{6.13}$$

$$C = \frac{\partial^2 F}{\partial b^2} = 2N, \tag{6.14}$$

$$D = AC - B^2 = 4N\left(\sum_{i=1}^{N} X_i{}^2\right) - 4\left(\sum_{i=1}^{N} X_i\right)^2. \tag{6.15}$$

Если *D*<0 функция $F(a,b)$ является гиперболическим параболоидом, а *M* – седловой точкой, в которой нет экстремума.

Если *D*>0 функция $F(a,b)$ является эллиптическим параболоидом. При этом если *A*>0, то в точке *M* наблюдается минимум, а иначе – максимум.

Кроме того вместо формул (6.9), (6.10) можно использовать следующие выражения для нахождения коэффициентов *a* и *b*:

$$a = \frac{\overline{XY} - \overline{X}\cdot\overline{Y}}{\overline{X^2} - \overline{X}^2}, \; b = \frac{\overline{Y}\,\overline{X^2} - \overline{X}\cdot\overline{XY}}{\overline{X^2} - \overline{X}^2}, \tag{6.16}$$

здесь $\overline{XY}$ – среднее выборки, содержащей поэлементные произведения векторов *X* и *Y*;

$\overline{X^2}$ - среднее выборки, содержащей поэлементные возведение значений вектора *X* в квадрат.

Таким образом, ***алгоритм нахождения функции одномерной линейной регрессии*** следующий:

1 Проверка исходных данных через коэффициент линейной корреляции или на диаграмме рассеяния на линейность

2 Выдвижение нулевой гипотезы о линейной зависимости отклика от фактора. H_0: $Y=f(X)=aX+b$, где *a* и *b* – неизвестные.

3 Построение функции суммарной ошибки $F(a,b)=\sum \varepsilon_i{}^2$, где $\varepsilon_i = Y_i - Y_i^T = Y_i - (aX_i + b)$.

4 Минимизация функции суммарной ошибки:

а) из необходимых условий экстремума находятся *a* и *b*;

б) с помощью достаточных условий проверяется, что при найденных *a* и *b* суммарная ошибка минимальна.

6.3 Проверка адекватности регрессионной модели

Для анализа адекватности регрессионной модели используют коэффициент детерминации R^2. Выборочный параметр коэффициента детерминации определяется формулой

$$R^2 = \frac{\sum (Y_i^T - \bar{Y})^2}{\sum (Y_i - \bar{Y})^2}. \qquad (6.17)$$

Коэффициент детерминации всегда находится в пределах интервала [0;1]. Если значение R^2 близко к единице, это означает, что построенная модель объясняет почти всю изменчивость переменных. Близкое к нулю значение R^2 означает плохое качество построенной модели.

Коэффициент детерминации показывает, на сколько процентов (R^2)·100% найденная функция регрессии описывает связь между исходными значениями *Y* и *X*. В соотношении (6.17)

$(Y_i^T - \bar{Y})$ – объясненное регрессионной моделью отклонение,

$(Y_i - \bar{Y})$ - общее отклонение.

Величина (1-R^2)·100% показывает, сколько процентов отклонения значений параметра *Y* обусловлены факторами, не включенными в регрессионную модель.

При высоком значении коэффициента детерминации ($R^2 \geq 75\%$) можно делать прогноз значения отклика для конкретного значения фактора в пределах диапазона исходных данных. При прогнозах значений, не входящих в диапазон исходных данных, справедливость полученной модели гарантировать нельзя. Это объясняется тем, что может проявиться влияние новых факторов, которые модель не учитывает.

Оценка значимости уравнения регрессии осуществляется с помощью статистической проверки гипотез и критерия Фишера по следующему алгоритму:

1 Нулевой гипотезой является предположение о неадекватности построенной модели, которая выражается в равенстве нулю генерального коэффициента детерминации. H_0: $R^2=0$.

Т.е. даже если выборочный коэффициент детерминации близок к 1, настоящий – генеральный, характеризующий генеральную совокупность – равен 0, а значение выборочного объясняется только неудачными данными в выборке.

2 Вычисляется статистика критерия по формуле

$$F_{\text{набл}} = \frac{R^2}{1-R^2} \cdot \frac{N-p-1}{p}. \qquad (6.18)$$

3 Задается уровень значимости α (обычно равный 0,5).

4 По статистическим таблицам значений распределения Фишера с числом степеней свободы $k_1=1$, $k_2=N-2$ (для линейной модели) для заданного уровня значимости α находится критическое значение $F_{\text{кр}}$.

5 Вывод о принятии или отклонении нулевой гипотезы делается путем сравнения статистики критерия (6.18) с критическим значением $F_{\text{кр}}$. Если выполняется неравенство

$$F_{\text{набл}} \le F_{\text{кр}}, \qquad (6.19)$$

нулевая гипотеза принимается, то есть признается статистическая незначимость или ненадежность уравнения регрессии. Иначе гипотеза отклоняется, следовательно, уравнение регрессии считается статистически значимым.

7 Корреляционный анализ

7.1 Задача корреляционного анализа

Как и в прошлой лекции будем рассматривать полученные в ходе эксперимента наборы данных – значения некоторых параметров. Данные значения записаны в виде векторов $X=\{x_1,x_2,\ldots,x_N\}$, $Y=\{y_1,y_2,\ldots,y_N\}$, и так далее.

Задачей корреляционного анализа является обнаружение взаимосвязи между двумя параметрами, которые можно рассматривать как случайные величины, и количественная оценка степени неслучайности их совместного изменения. Исследуемые величины могут быть как двумя разными показателями в одной выборке, так и двумя различными выборками. Например, значения роста и веса опрошенных людей (два параметра одной выборки) или коэффициент усидчивости у каждого близнеца из опрошенных пар близнецов (две различные выборки).

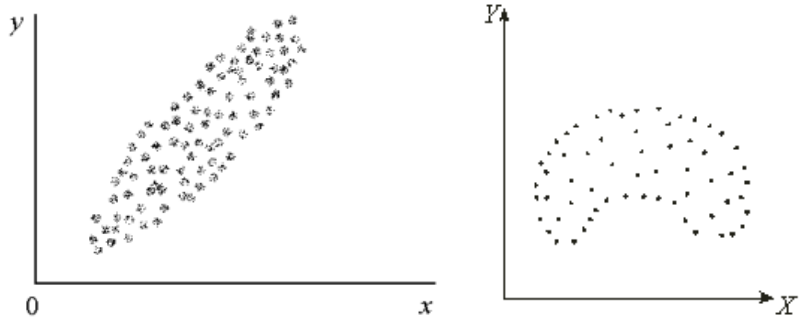
Если связь между величинами существует, корреляционный анализ показывает, сопровождается ли увеличение одного показателя увеличением или уменьшением другого и насколько сильно изменение одного показателя влияет на изменение другого.

Таким образом, корреляционный анализ помогает установить возможность предсказания вероятных значений одного показателя с помощью известных значений другого.

Наглядным изображением исходных данных корреляционного анализа служит корреляционное поле - график, где на оси абсцисс откладывается шкала для одного показателя (или выборки), по оси ординат – для другого, при этом на поле графика точками отмечаются значения исходных данных. По расположению точек можно судить о наличии или отсутствии связи, ее силе и характере влияния изменения одной переменной на изменение другой.

7.2 Классификация корреляционной связи

Связь между величинами может быть как линейной, так и нелинейной. Вид корреляционного поля в этих случаях приведен на рисунке 7.1.



а) б)

Рисунок 7.1 – Корреляционное поле в случае а) линейной и б) нелинейной связи между величинами

По силе корреляционные связи делятся на

- функциональную – есть жесткая зависимость между двумя параметрами, которую можно записать в виде функции без сглаживания;

- сильную;
- умеренную;
- слабую;
- отсутствующую – связи нет.

По направлению корреляционные связи делятся на

- положительные, характеризующие прямую зависимость между параметрами, когда увеличение одного параметра приводит к увеличению другого;
- отрицательные, характеризующие обратную зависимость между параметрами, когда увеличение одного параметра приводит к уменьшению другого.

Корреляционные поля для различных видов корреляционной связи по силе и направлению приведены на рисунке 7.2.

Для определения силы и направления связи между параметрами используется понятие ***коэффициента корреляции***. Коэффициент корреляции рассчитывается только для случая линейной взаимосвязи между параметрами. Рассчитанный коэффициент корреляции для нелинейной закономерности может привести исследователя в заблуждение, показав либо отсутствие связи, либо менее сильную связь.

Коэффициент линейной корреляции для двух случайных векторов $X=\{x_i\}$, $Y=\{y_i\}$, $i=1\ldots N$, рассчитывается по формуле

$$r = \frac{\overline{XY} - \overline{X}\,\overline{Y}}{\sigma_X \sigma_Y}, \qquad (7.1)$$

где $\overline{XY}$ - среднее значение выборки, состоящей из поэлементного умножения элементов исходных выборок, $\overline{XY} = \frac{1}{N}\sum_{i=1}^{N} x_i y_i$;

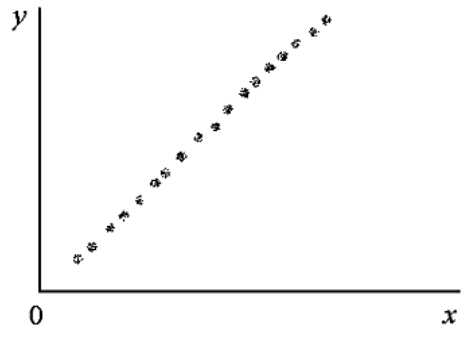
$\overline{X}$, $\overline{Y}$ - средние значения выборок X и Y соответственно,

$$\overline{X} = \frac{1}{N}\sum_{i=1}^{N} x_i , \qquad (7.2)$$

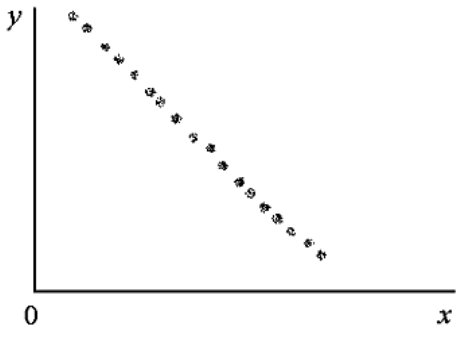
$$\overline{Y} = \frac{1}{N}\sum_{i=1}^{N} y_i ; \qquad (7.3)$$

σ_X , σ_Y - среднеквадратические отклонения (СКО) выборок X и Y соответственно,

$$\sigma_X = \sqrt{\frac{\sum_{i=1}^{N}(x_i - \overline{X})^2}{N}} ,. \qquad (7.4)$$



а)



б)

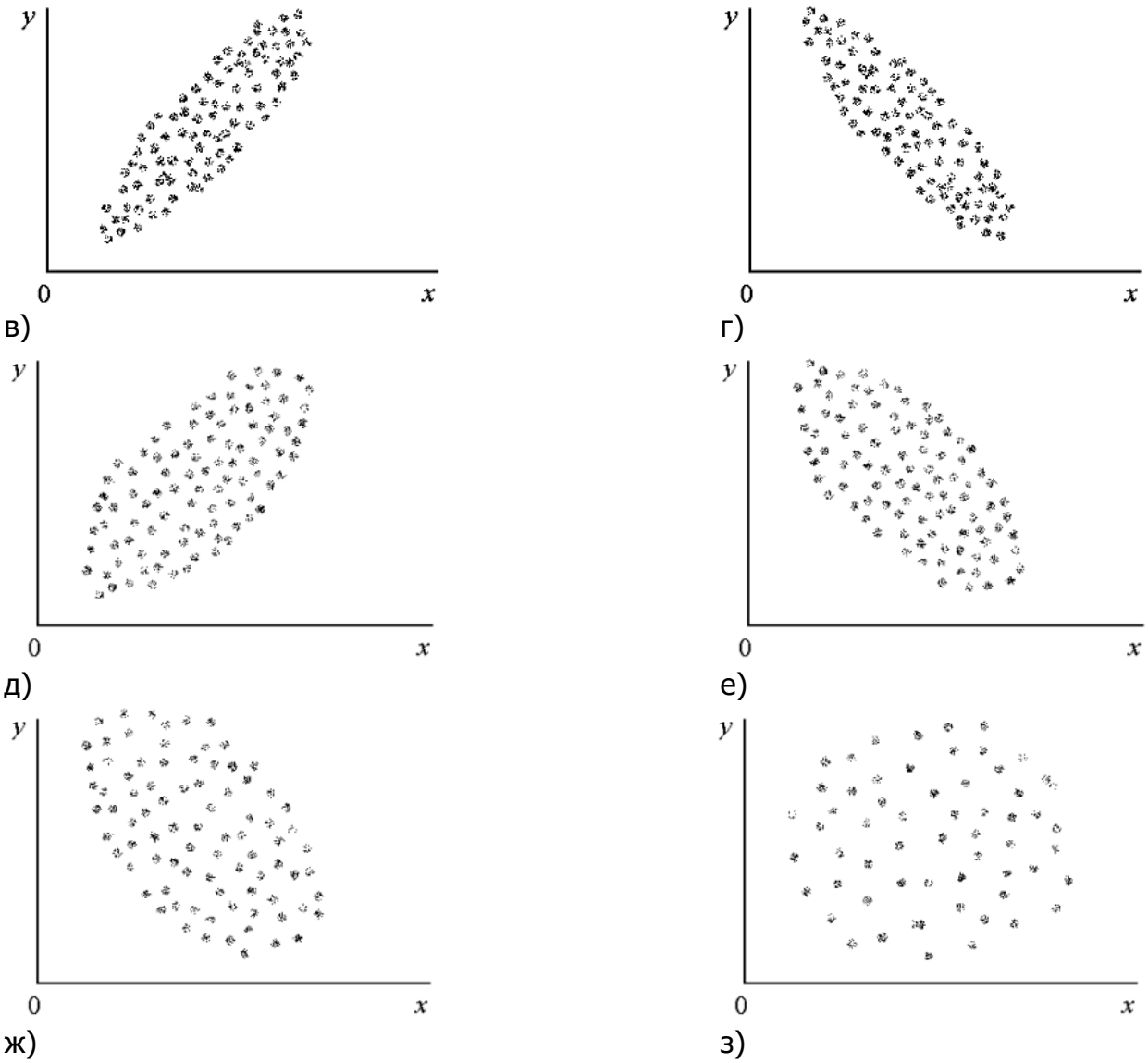


Рисунок 7.2 – Виды корреляционной связи на корреляционном поле
а) функциональная положительная; б) функциональная отрицательная;
в) сильная положительная; г) сильная отрицательная;
д) умеренная положительная; е) умеренная отрицательная;
ж) слабая отрицательная; з) связь отсутствует

$$\sigma_Y = \sqrt{\frac{\sum_{i=1}^{N}(y_i - \overline{Y})^2}{N}}\,. \qquad (7.5)$$

Формула (7.1) может быть записана также в следующем виде:

$$r = \frac{\sum_{i=1}^{N}(x_i - \overline{X})(y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{N}(x_i - \overline{X})^2}\sqrt{\sum_{i=1}^{N}(y_i - \overline{Y})^2}}\,, \qquad (7.6)$$

Формулы (7.1) и (7.6) тождественны.

Для малых объемов выборки ($N \leq 100$) производится корректировка коэффициента корреляции по формуле

$$r' = r\left(1 + \frac{1 - r^2}{2(N-3)}\right). \qquad (7.7)$$

Коэффициент корреляции r принимает значения от -1 до 1 включительно. Его знак говорит о характере связи (положительная или отрицательная), а модуль – о силе связи.

В случае, когда коэффициент корреляции равен нулю, корреляция между величинами X и Y отсутствует, то есть изменение X не приводит к изменению Y. Вид корреляционного поля при $r = 0$ приведен на рисунке 7.2 з).

При $|r| = 1$ наблюдается строгая функциональная зависимость между переменными, то есть существует такая функция $y = f(x)$, которая описывает изменение значения величины Y при изменении значения величины X. Вид корреляционного поля при вырождении корреляционной зависимости в функциональную для положительной и отрицательной связей приведен на рисунках 7.2 а) и б).

По мере уменьшения модуля коэффициента корреляции до нуля зависимость одной переменной от другой все больше уменьшается, то есть «облако» значений на корреляционной плоскости становится шире и все более округлым. В случае, когда коэффициент корреляции по модулю равен единице, «облако» значений «концентрируется» в график функции зависимости.

Сила связи между параметрами в зависимости от модуля коэффициента корреляции приведена в таблице 7.1. Вид данных на корреляционном поле для сильной, умеренной и слабой связей приведен на рисунках 7.3 в)-ж).

Таблица 7.1 – Сила связи между параметрами

Значение r	Сила связи
$\|r\| = 1$	функциональная
$0{,}7 \le \|r\| < 1$	сильная
$0{,}5 \le \|r\| \le 0{,}7$	умеренная
$0{,}3 \le \|r\| \le 0{,}5$	слабая
$0 < \|r\| \le 0{,}3$	практически отсутствует
$\|r\| = 0$	отсутствует

7.3 Связь с регрессионным анализом

Корреляционный анализ связан с регрессионным. Корреляционный анализ устанавливает силу и направление связи переменных, степень неслучайности их изменения, регрессионный анализ позволяет записать сглаженную функцию их взаимосвязи в диапазоне изменения данных. Таким образом, после применения корреляционного и регрессионного анализов в интервале изменения данных можно предсказать значения отклика по значениям фактора.

Понятие корреляционного поля в корреляционном анализе равносильно понятию диаграммы рассеяния в регрессионном.

Кроме того, между линейной корреляцией и линейным одномерным регрессионным анализом существует следующая связь:

Если коэффициент линейной корреляции известен, то уравнение линейной регрессии случайного вектора Y на случайный вектор X имеет вид:

$$Y - \overline{Y} = r\frac{\sigma_Y}{\sigma_X}\left(X - \overline{X}\right). \qquad (7.8)$$

Уравнение линейной регрессии X на Y может быть записано в виде:

$$X - \overline{X} = r\frac{\sigma_X}{\sigma_Y}\left(Y - \overline{Y}\right). \quad (7.9)$$

В случае записи уравнения регрессии в виде $y = ax + b$ формула (7.8) преобразуется к виду

$$Y = r\frac{\sigma_Y}{\sigma_X}X + \left(\overline{Y} - r\frac{\sigma_Y}{\sigma_X}\overline{X}\right). \quad (7.10)$$

Таким образом, можно получить следующие значения параметров функции линейной регрессии:

$$a = r\frac{\sigma_Y}{\sigma_X}, \quad (7.11)$$

$$b = \overline{Y} - r\frac{\sigma_Y}{\sigma_X}\overline{X}. \quad (7.12)$$

7.4 Проверка статистических гипотез в корреляционном анализе

В корреляционном анализе посредством проверки статистических гипотез проверяются два основных предположения:

- значимость генерального коэффициента линейной корреляции;
- значимость различия между двумя коэффициентами корреляции

Выборочный коэффициент корреляции r – коэффициент, рассчитанный по данным выборок X и Y - является оценкой генерального коэффициента корреляции, который показывает реальную связь между параметрами X и Y. Из-за конечного размера выборок возможен случай, когда рассчитанный по выборкам r будет близок к единице, а коэффициент корреляции генеральной совокупности будет нулевым. Т.е. выборочный коэффициент корреляции покажет отсутствующую (нулевую) на генеральной совокупности сильную связь между параметрами.

Для доказательства ***значимости генерального коэффициента корреляции*** применяемся следующий алгоритм.

1 Выдвигаются нулевая и альтернативная гипотезы:

- нулевая - о равенстве нулю генерального коэффициента корреляции: $H_0 : r_S = 0$;
- альтернативная - $H_1 : r_S \neq 0$

Если нулевая гипотеза отвергается, принимается альтернативная.

2 Задается уровень значимости α, который соответствует вероятности совершить ошибку первого рода. Уровень значимости обычно считается равным 0,05 или 0,01.

3 Для большого объема выборки ($N \geq 100$) вычисляется статистика

$$t_{\text{набл}} = \frac{|r|}{\sqrt{1-r^2}}\sqrt{N-2}, \quad (7.13)$$

а для малого –

$$t_{\text{набл}} = 0{,}5\ln\left(\frac{1+r'}{1-r'}\right)\sqrt{N-3}, \quad (7.14)$$

где r' - скорректированное по формуле (7.7) значение коэффициента корреляции.

4 Вычисленная статистика $t_{\text{набл}}$ сравнивается с критическим значением $t_{\text{кр}}$ - табличным значением распределения Стьюдента $t(P=1-\alpha,\ \infty)$, равным 1,96 для α=0,05 и 2,576 для α=0,01. Если

$$t_{\text{набл}} > t_{\text{кр}}, \tag{7.15}$$

то H_0 отвергается, а следовательно, генеральный коэффициент корреляции значимо больше нуля.

Значение коэффициента корреляции может меняться в зависимости от размера выборки. С помощью второго алгоритма статистической проверки гипотез можно оценить, принадлежат ли две пары выборок одной генеральной совокупности.

Пусть имеются две пары выборок $X1=\{x1_i\}$, $Y1=\{y1_i\}$, $i=1\ldots N$, и $X2=\{x2_j\}$, $Y2=\{y2_j\}$, $j=1\ldots M$, $M\neq N$. Для каждой из этих пар рассчитан коэффициент корреляции $r1$ и $r2$ соответственно, причем $r1\neq r2$. Необходимо выяснить, взяты ли обе эти пары из общей генеральной совокупности, то есть имеют ли они общий генеральный коэффициент корреляции.

Для решения этой задачи проверяется ***гипотеза о значимости различия между двумя коэффициентами корреляции*** по следующему алгоритму.

1 Выдвигается нулевая и альтернативная гипотезы:

- нулевая - о незначимости различий между двумя генеральными коэффициентами линейной корреляции $H_0: r1_S = r2_S = r_S$;
- альтернативная - $H_1: r1_S \neq r2_S$.

Если нулевая гипотеза отвергается, принимается альтернативная.

2 Задается уровень значимости α (обычно 0,05 или 0,01).

3 Вычисляется статистика

$$t_{\text{набл}} = 0{,}5\ln\left(\frac{(1+r1)(1-r2)}{(1-r1)(1+r2)}\right)\frac{1}{\sqrt{\frac{1}{N-3}+\frac{1}{M-3}}}. \tag{7.16}$$

4 Вычисленная статистика $t_{\text{набл}}$ также сравнивается с критическим значением $t_{\text{кр}}$ - табличным значением распределения Стьюдента $t(P=1-\alpha,\ \infty)$. Аналогично, если

$$t_{\text{набл}} > t_{\text{кр}}, \tag{7.17}$$

то H_0 отвергается, следовательно, нельзя считать, что обе пары взяты из одной генеральной совокупности.

8 Дисперсионный анализ

8.1 Основные понятия дисперсионного анализа

Дисперсионный анализ (*ANOVA, Analysis of variances*) применяется к результатам наблюдений, которые зависят от различных одновременно действующих факторов, для выбора наиболее значимых факторов и оценки их влияния на исследуемый процесс. Например, дисперсионный анализ применяется для оценки влияния режима нагрузки на долговечность технического изделия.

Влияние различных факторов на исследуемые случайные величины (например, влияние технологического способа изготовления или режима нагрузки на долговечность технического изделия) приводит к изменению значений параметров распределения вероятностей этих величин — первого момента (среднего), второго момента (дисперсии) или моментов более высокого порядка.

Суть ***дисперсионного анализа*** заключается в разделении общей дисперсии изучаемого признака на отдельные компоненты, обусловленные влиянием конкретных факторов, и проверке гипотез о значимости влияния этих факторов на среднее значение наблюдаемой случайной величины.

Классические методы дисперсионного анализа основываются на следующих предпосылках:

- распределение исходных случайных величин нормально;
- дисперсии экспериментальных данных одинаковы для всех условий эксперимента (т. е. для экспериментов, выполненных на различных уровнях изучаемого фактора).

Поэтому проведение дисперсионного анализа предваряет проверка нормальности распределения изучаемой случайной величины, и неразличимость дисперсий изучаемых совокупностей, например, с помощью критерия Кохрена, Бартлетта и т.д..

Исходным материалом для дисперсионного анализа служат данные исследования трех и более выборок: $x_1,...,x_m$, которые могут быть как равными, так и неравными по численности, как зависимыми, так и независимыми.

Набор значений откликов, полученных при фиксированных уровнях факторов, называется ***группой*** (т.о. это набор данных, объединенных по одному признаку). Изменение откликов в дисперсионном анализе называется ***градацией***, при этом различают

- ***межгрупповую градацию*** *– изменение откликов, соответствующее уровням факторов;*
- ***внутригрупповую градацию*** *– изменение откликов внутри одной выборки, соответствующей одному уровню факторов.*

Основные понятия дисперсионного анализа представлены на рисунке 8.1.

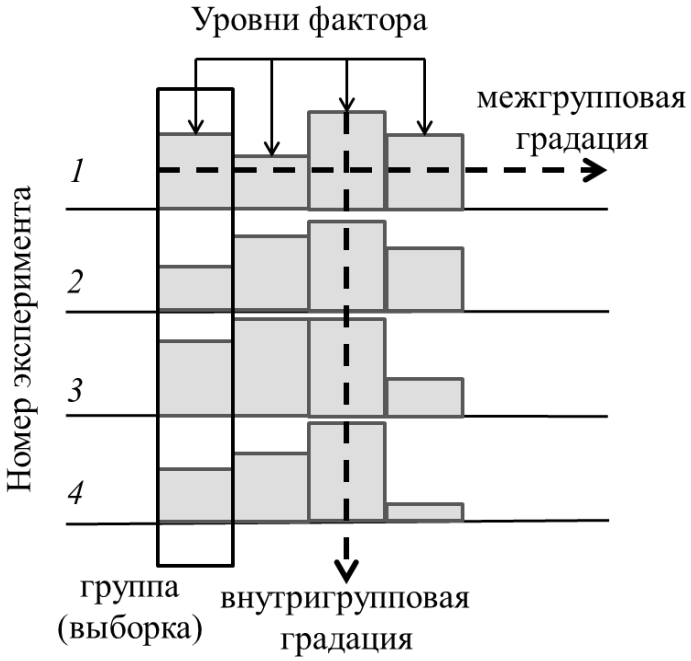


Рисунок 8.1 – Основные понятия дисперсионного анализа

В зависимости от количества факторов и откликов различают:

- однофакторный и многофакторный дисперсионный анализ, который исследует воздействие одного или одновременно действующих нескольких независимых переменных;
- одномерный и многомерный дисперсионный анализ, который рассматривает одну или несколько зависимых переменных.

Кроме того дисперсионный анализ делится в зависимости от типа выборки на две следующие разновидности:

- анализ несвязных (различных) выборок, когда исследование проводится в разных группах при разных уровнях фактора;
- анализ связных выборок, когда исследование проводится в одной и той же группе при разных уровнях фактора.

Дерево классификаций дисперсионного анализа приведено на рисунке 8.2.

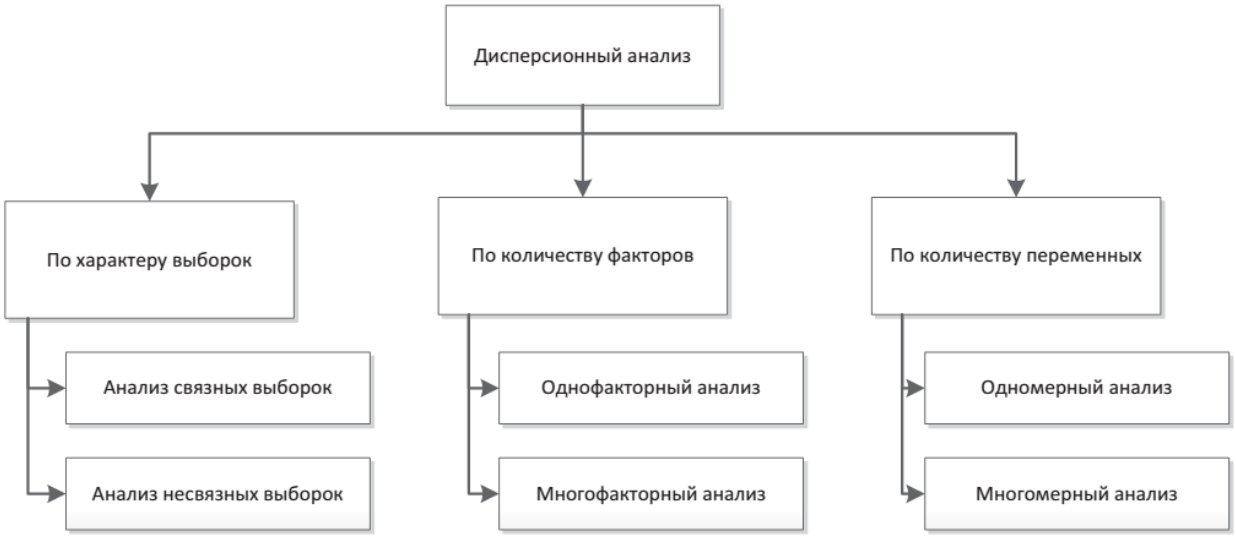


Рисунок 8.2 - Классификация дисперсионного анализа

8.2 Методы предварительной оценки данных

Как было сказано выше, перед применением дисперсионного анализа необходимо убедиться, что

- распределение исходных случайных величин внутри каждой выборки нормально;
- дисперсии экспериментальных данных одинаковы для всех условий эксперимента (т. е. для экспериментов, выполненных на различных уровнях изучаемого фактора).

Для доказательства нормальности распределения выборок можно воспользоваться любым из описанных в лекциях 4-5 способом, либо узкоспециализированными критериями нормальности.

Для сравнения нескольких дисперсий существует довольно много критериев, например, критерий Бартлетта, критерий Нейманна-Пирсона. Рассмотрим ***критерий Кохрена***, который применяется в случае выборок равных объемов.

Пусть $s_1^2, s_2^2, ..s_m^2$ - взаимно независимые выборочные оценки дисперсий $\sigma_1^2, \sigma_2^2, ..\sigma_m^2$ по выборкам объема n.

1 Выдвигается нулевая гипотеза о незначимости отличий между генеральными дисперсиями H_0: $\sigma_1^2 = \sigma_2^2 = ... = \sigma_m^2$.

2 По имеющимся значениям оценок дисперсий вычисляется статистика

$$g_{\text{набл}} = \frac{\max\limits_{1 \le i \le m} s_i^2}{\sum\limits_{i=1}^{m} s_i^2}. \qquad ((8.1)$$

3 Задается уровень значимости α (обычно равный 0,05).

4 Рассчитывается критическое значение

$$g_{\text{кр}} = \frac{F_{\frac{m+1-\alpha}{m}}\left(n-1;(n-1)(m-1)\right)}{m-1+F_{\frac{m-1+\alpha}{m}}\left(n-1;(n-1)(m-1)\right)}, \qquad (8.2)$$

где $F_x(f_1,f_2)$ – x-квантиль распределения Фишера с f_1 и f_2 степенями свободы.

Значения распределения Фишера заданы таблично. Кроме того, имеются таблицы значений $g_{\text{кр}}$ для различных уровней значимости в зависимости от n и m, таким образом формулой (8.2) пользоваться необязательно.

5 Если выполняется условие

$$g_{\text{набл}} \le g_{\text{кр}}. \qquad (8.3)$$

нулевая гипотеза принимается и генеральные дисперсии считаются равными.

8.3 Одномерный однофакторный дисперсионный анализ

Одномерный однофакторный дисперсионный анализ позволяет оценить влияние одного фактора на один отклик.

Пусть у нас есть m выборок $x_1,...,x_m$ одинакового объема n.

Исходные данные одномерного однофакторного дисперсионного анализа могут быть представлены в виде так называемой статистической таблицы, образец которой приведен в таблице 8.1.

Таблица 8.1 – Статистическая таблица для одномерного однофакторного дисперсионного анализа

Номер эксперимента	Уровни фактора *Fact*				
	F_1	…	F_j	…	F_m
1	x_{11}	…	x_{1j}	…	x_{1m}
…	…		…		…
i	x_{i1}	…	x_{ij}	…	x_{im}
…	…		…		…
n	x_{n1}	…	x_{nj}	…	x_{nm}

Одномерный однофакторный дисперсионный анализ представляет собой проверку статистической гипотезы о равенстве средних значений групп. Если средние групп равны, то фактор не имеет влияния на отклик, иначе – влияние фактора существенно.

В процессе анализа проводится расчет трех видов дисперсий:

- общей (дисперсия комплекса);
- межгрупповой (факторная);
- внутригрупповой (остаточная).

Ниже описан ***алгоритм проведения дисперсионного анализа***.

Перед началом дисперсионного анализома задается уровень значимости α (обычно 0,05).

1 Выдвигается нулевая гипотеза о равенстве средних значений групп $H_0 : Mx_1 = Mx_2 = Mx_3 = \ldots = Mx_m$.

2 Находятся три вида средних значений:

- внутригрупповое среднее $\bar{x}_{*j}$

$$\bar{x}_{*j} = \frac{1}{n}\sum_{i=1}^{n} x_{i,j}\ ; \qquad (8.4)$$

- межгрупповое среднее $\bar{x}_{i*}$

$$\bar{x}_{i*} = \frac{1}{m}\sum_{j=1}^{m} x_{i,j}\ ; \qquad (8.5)$$

- общее среднее

$$\bar{x}_{**} = \frac{1}{nm}\sum_{i=1}^{n}\sum_{j=1}^{m} x_{i,j}\ . \qquad (8.6)$$

Средние заносятся в статистическую таблицу так, как показано на рисунке 8.3.

Номер эксперимента	Уровни фактора *Fact*					
	F_1	…	F_j	…	F_m	Σ/m
1	x_{11}	…	x_{1j}	…	x_{1m}	$\bar{x}_{1*}$
…	…		…		…	…
i	x_{i1}	…	x_{ij}	…	x_{im}	$\bar{x}_{i*}$
…	…		…		…	…
n	x_{n1}	…	x_{nj}	…	x_{nm}	$\bar{x}_{n*}$
Σ/n	$\bar{x}_{*1}$		$\bar{x}_{*j}$		$\bar{x}_{*m}$	$\bar{x}_{**}$ общее среднее

межгрупповые средние

внутригрупповые средние

Рисунок 8.3 – Средние в статистической таблице

3 Проводится расчет трех видов сумм квадратов отклонений:

- общей суммы квадратов отклонений от общего среднего

$$R_{\text{общ}} = \sum_{i=1}^{n}\sum_{j=1}^{m}\left(x_{ij} - \bar{x}_{**}\right)^2; \quad (8.7)$$

- факторной суммы квадратов отклонений групповых средних от общего среднего, которая характеризует влияние фактора

$$R_{\text{факт}} = n\sum_{j=1}^{m}\left(\bar{x}_{*j} - \bar{x}_{**}\right)^2. \quad (8.8)$$

- остаточной суммы квадратов отклонений, характеризующей внутригрупповое рассеяние

$$R_{\text{ост}} = R_{\text{общ}} - R_{\text{факт}}. \quad (8.9)$$

Внутригрупповое рассеяние не может быть предсказано или объяснено. Факторная сумма квадратов отклонений объясняется различиями между средними значениями в группах, то есть рассеянием между группами.

4 Проводится расчет трех видов несмещенных выборочных дисперсий:

- общей

$$S^2_{\text{общ}} = \frac{R_{\text{общ}}}{nm-1}, \quad (8.10)$$

- факторной, характеризующей систематическую (межгрупповую) дисперсию

$$S^2_{\text{факт}} = \frac{R_{\text{факт}}}{m-1}, \quad (8.11)$$

- остаточной, характеризующей случайную (внутригрупповую) дисперсию

$$S^2_{\text{ост}} = \frac{R_{\text{ост}}}{m(n-1)}. \quad (8.12)$$

5 Проводится расчет статистики – соотношения систематической дисперсии к случайной - по формуле

$$F_{\text{набл}} = \frac{S^2_{\text{факт}}}{S^2_{\text{ост}}}. \quad (8.14)$$

Данная статистика используется, поскольку отношение двух выборочных дисперсий распределено по закону Фишера.

6 По числу степеней свободы $f_1=m-1$, $f_2=m(n-1)$ и уровню значимости α находится $F_{\text{кр}}$ - значения α-квантиля распределения Фишера в критической точке, , по числу степеней свободы Данная величина задана таблично.

7 Если

$$F_{\text{набл}} > F_{\text{кр}}, \quad (8.15)$$

то нулевая гипотеза отвергается. Это значит, что фактор *Fact* оказывает существенное воздействие и его надо учитывать.

В случае принятия нулевой гипотезы фактор оказывает незначительное влияние и им можно пренебречь.

Иногда дисперсионный анализ применяется для того, чтобы установить, что выборки однородны: дисперсии выборок одинаковы по предположению, а анализ покажет, равны ли математические ожидания. Если да, то выборки можно объединить в одну и, исследовав ее, получить более полную информацию.

8.4 Многофакторный дисперсионный анализ

Многофакторный дисперсионный анализ обладает большей информативностью по сравнению с однофакторным анализом, однако чем больше факторов рассматривается, тем сложнее проведение анализа. Ограничимся рассмотрением двухфакторного одномерного анализа.

Пусть на параметр воздействуют два фактора, *A* и *B*. В ходе двухфакторного дисперсионного анализа можно проверить следующие гипотезы:

- равенство средних под влиянием фактора *A*;
- равенство средних под влиянием фактора *B*;
- отсутствие взаимодействия факторов *A* и *B*.

Рассмотрим доказательство первых двух гипотез.

Статистическая таблица для двухфакторного дисперсионного анализа представлена на рисунке 8.4.

Для проведения анализа вычисляются суммы:

$$R_1 = \sum_{i=1}^{k}\sum_{j=1}^{m} x_{ij}^2; \quad (8.16)$$

$$R_2 = \frac{1}{m}\sum_{i=1}^{k} X_i^2; \quad (8.17)$$

$$R_3 = \frac{1}{k}\sum_{j=1}^{m} X_{j'}^2; \quad (8.18)$$

$$R_4 = \frac{1}{mk}\left(\sum_{j=1}^{m} X_{j'}\right)^2 = \frac{1}{mk}\left(\sum_{i=1}^{k} X_i\right)^2. \quad (8.19)$$

B	A						Σ
	A_1	A_2	…	A_i	…	A_k	
B_1	x_{11}	x_{21}	…	x_{i1}	…	x_{k1}	X_1`
B_2	x_{12}	x_{22}	…	x_{i2}	…	x_{k2}	X_2`
…	…	…	…	…	…	…	…
B_j	x_{1j}	x_{2j}	…	x_{ij}	…	x_{kj}	X_j`
…	…	…	…	…	…	…	…
B_m	x_{1m}	x_{2m}	…	x_{im}	…	x_{km}	X_m`
Σ	X_1	X_2	…	X_i	…	X_k	

Рисунок 8.4 – Средние в статистической таблице

После этого находятся оценки дисперсий:

$$S^2_{\text{общ}} = \frac{R_1 + R_4 - R_2 - R_3}{(k-1)(m-1)}; \quad (8.20)$$

$$S^2_{\text{A}} = \frac{R_2 - R_4}{k-1}; \quad (8.21)$$

$$S^2_{\text{B}} = \frac{R_3 - R_4}{m-1}; \quad (8.22)$$

Если

$$\frac{S^2_{\text{A}}}{S^2_{\text{общ}}} > F_\alpha\left(k-1,\ (k-1)(m-1)\right), \quad (8.23)$$

влияние фактора *A* с достоверностью α принимается значимым.

Если

$$\frac{S^2_{\text{B}}}{S^2_{\text{общ}}} > F_\alpha\left(m-1,\ (k-1)(m-1)\right), \quad (8.24)$$

влияние фактора *B* с достоверностью α принимается значимым.

Для проверки отсутствие взаимодействия факторов *A* и *B* необходимо, чтобы в каждой клетке таблицы (т.е. при каждом сочетании факторов *A* и *B*) было не одно, а серия наблюдений.

9 Способы получения экспертных оценок

9.1 Общие сведения о методах экспертной оценки

Проведение экспертной оценки необходимо для выбора одного или нескольких проектов из предложенного множества, а также для расстановки проектов в ряд от лучшего к худшему. Экспертную оценку проводит коллектив экспертов.

Пусть коллектив экспертов состоит из M человек, а количество проектов, предложенных к рассмотрению, S. Условием проведения экспертизы является соблюдение неравенства $M>S$, т.е. превалирование числа экспертов над числом проектов.

В процессе работы эксперты анализируют проекты и выставляют каждому проекту свою собственную оценку, в результате чего формируется так называемая матрица оценок (рисунок 9.1). Далее с помощью одного или нескольких экспертных методов по матрице оценок

- либо выбирается проект-победитель (система голосования с единственным победителем, *single-winner*);
- либо определяются несколько проектов, занявших призовые места с указанием порядка мест (система голосования с множественными победителями, *multiple-winner*).

Кроме самих результатов экспертизы теория рассматривает и оценку экспертов по результатам их предпочтений, а также установление степени согласованности мнений экспертной комиссии.

Рассмотрим два класса методов экспертизы: голосование и ранжирование.

Эксперты	Проекты			
	1	2	…	S
1	r_{11}	r_{12}	…	r_{1S}
2	r_{21}	r_{22}	…	r_{2S}
…	…	…	…	…
M	r_{M1}	r_{M2}	…	r_{MS}

Рисунок 9.1 – Матрица оценок при экспертизе проектов

9.2 Методы голосования

Голосование представляет собой отдание экспертом своего голоса за один из предложенных проектов, таким образом, матрица оценок может быть представлена в виде нулей и единиц, причем в каждой строке возможна лишь одна единица, соответствующая голосу «за».

Известны следующие методы голосования:

- голосование большинством (правило относительного большинства);
- голосование большинством 2/3;
- голосование с правом вето;
- талонное голосование;

- туровое голосование.

Голосование большинством голосов - самый простой способ голосования: победителем объявляется проект, набравший абсолютное большинство голосов. Возможно, принятое решение будет неоптимальным за счет низкой квалификации (другой узкой специализации) экспертов, что является недостатком данного метода.

Голосование большинством можно улучшить наложением ограничений. В зависимости от типа ограничения, голосование большинством преобразуется в другие методы голосования.

При ***голосовании большинством 2/3*** победителем назначается тот проект, за который проголосовали 2/3 или более экспертов. Это повышает порог принятия решения и применяется при принятии ответственных решений.

Данный метод может приводить к тупиковой ситуации, когда ни один проект не будет поддержан 2/3 голосов.

При ***голосовании с правом вето*** любой эксперт имеет право заблокировать общее намечающееся решение, если считает его невыгодным или опасным. Данная процедура является еще более жесткой и применяется только в самых ответственных ситуациях. Способ также может приводить к тупиковой ситуации.

Талонное голосование может быть реализовано, если одна и та же группа экспертов проводит несколько голосований (*N)* в некоторый отрезок времени. Голосовать можно только при наличии талона. Перед первым голосованием из серии каждому эксперту выдается *L* талонов, причем $L<N$. Таким образом, регламентируется возможность эксперту пропустить голосование, если он не уверен в своей квалификации по данному вопросу, т.е. эксперт оценивает не только проекты, но и себя.

Туровое голосование проводится в несколько этапов (туров) и может делиться на следующие методы:

- голосование в два тура, когда проводится обычный тур голосования. Если какой-либо проект набирает строгое большинство голосов, то он и объявляется победителем. В противном случае проводится второй тур голосования по правилу большинства для двух-трех проектов, набравших наибольшее число голосов в первом туре;
- голосование с последовательным убыванием проводится в общем случае в несколько туров. Если какой-либо проект набирает строгое большинство голосов, то он объявляется победителем. В противном случае проводится второй тур голосования, но без участия проекта, набравшего минимальное количество голосов в первом туре. Затем процесс повторяется, пока не определится проект-победитель. При *n* рассматриваемых проектах может потребоваться *n*-1 туров;
- голосование без отсеивания проектов предполагает улучшение результатов оценки через ознакомление экспертов с результатами каждого тура. В результате такой процедуры эксперты, видя и оценивая общий результат, могут менять собственные предпочтения, исключая случайность и приходя к равновесному мнению. Так работает, например, «метод Дельфы».

9.3 Методы ранжирования

Ранжирование - процедура упорядочения любых объектов по возрастанию или убыванию некоторого их свойства (разумеется, объекты должны обладать

этим свойством). Таким образом, методы ранжирования построены на введении каждым экспертом отношения порядка на множестве проектов *S*, причем отношение порядка может быть задано либо с помощью бальной шкалы (рангов), либо отношением «хуже-лучше». В последнем случае предпочтение экспертом проекта *x* проекту *y* обозначается $x \succ y$, а совокупность индивидуальных порядков предпочтений всех экспертов называется профилем предпочтений.

В случае введения отношения порядка с помощью рангов сумма рангов, поставленных экспертом по всем проектам, должна совпадать и быть равной

$$\sum_{j=1}^{S} r_{ij} = \frac{S(S+1)}{2}, \text{ для } i=1...M \qquad (9.1)$$

где r_{ij} – ранг, поставленный *i*-м экспертом *j*-му проекту.

Если эксперт присваивает двум или более проектам одинаковые ранги (т.е. считает эти проекты равнозначными), равенство (9.1) не выполняется, а следовательно, невозможно провести определение победителя. Тогда проводится пересчет таких рангов в т.н. стандартизированные ранги делением суммы мест, занимаемых связными рангами, на их число. То есть, если у *i*-го эксперта $r_{i1}=r_{i2}$ и эти проекты делят между собой места *a* и *b*, тогда данным проектам присваивается стандартизированный ранг *c*, определяемый по формуле

$$c = \frac{a+b}{2}. \qquad (9.2)$$

Решения, полученные разными методами, могут отличаться друг от друга. В ряде случаев отдельные методы могут не сработать, поэтому экспертизу проводят несколькими методами, а затем выводят общий ответ.

Рассмотрим некоторые из методов ранжирования.

Метод «Медиана Кемени» (*Kemeny*) основан на вычислении расстояния l_{ij} между мнениями *i*-го и *j*-го экспертов. Поскольку мнение эксперта - вектор значений, то расстояние между мнениями экспертов представляет собой расстояние между векторами и может быть вычислено с использованием различных вариантов расчета норм, например, по манхэттенской норме:

$$l_{ij} = \left|r_{i1} - r_{j1}\right| + \left|r_{i2} - r_{j2}\right| + ... + \left|r_{iS} - r_{jS}\right|. \qquad (9.3)$$

Значения для каждой пары экспертов заносятся в т.н. матрицу расстояний, которая в общем виде может быть представлена в следующем виде, приведенном на рисунке 9.2.

Матрица расстояний имеет размерность *M*x*M*, причем по главной диагонали расположены нули (т.к. не существует расхождения между мнением самого эксперта) Кроме того, согласно формуле (9.3) матрица симметрична относительно главной диагонали, т.е. $l_{ij}=l_{ji}$ для любых $i, j=1...M$.

	Эксперт 1	Эксперт 2	…	Эксперт *M*
Эксперт 1	0	l_{12}	…	l_{1M}
Эксперт 2	l_{21}	0	…	l_{2M}
…	…	…	…	…
Эксперт *M*	l_{M1}	l_{M2}	…	0

Рисунок 9.2 – Матрица расстояний

После заполнения матрицы расстояний вычисляются значения L_i - расстояния от i-го эксперта до всех остальных, после чего ищется минимальное расстояние L_i. Мнение соответствующего i-го эксперта является самым средним и объявляется результатом экспертизы. Процесс расчета расстояний между мнениями экспертов по матрице расстояний приведен на рисунке 9.3.

	Эксперт 1	Эксперт 2	…	Эксперт M	L_i
Эксперт 1	0	l_{12}	…	l_{1M}	$L_1=0+l_{12}+\ldots+l_{1M}$
Эксперт 2	l_{21}	0	…	l_{2M}	$L_2=l_{21}+0+\ldots+l_{2M}$
…	…	…	…	…	…
Эксперт M	l_{M1}	l_{M2}	…	0	$L_M=l_{M1}+l_{M2}+\ldots+0$
					$\min(L_1, L_2,\ldots L_M)$

Рисунок 9.3 – Расчет расстояний между мнениями экспертов

Геометрическая интерпретация показывает, что мнение выбранного эксперта находится примерно в центре многогранника, образованного вершинами, каждая из которых представляет собой мнение отдельного эксперта, а ребра — расстояния между мнениями экспертов. Самая удаленная от центра вершина, то есть эксперт с наибольшим значением L_i имеет самое ненадежное мнение. Данная величина может характеризовать квалификацию самого эксперта.

При проведении экспертной оценки ***методом большинства*** на первое место ставится тот проект, которому большинство экспертов присвоило первое место, на второе место – проект, которому большинство экспертов присвоило второе место и так далее.

В случае $S \geq 3$ при экспертизе данным методом могут возникнуть сомнения в справедливости результатов.

Метод Борда представляет собой систему голосования, предложенную в 1770 году Ж.-Ш. де Борда (*Borda*) и широко используется в современных условиях с целью тщательного учета предпочтений экспертов в условиях множества проектов.

Суть метода заключается в ранжировании экспертом всех проектов строго по убыванию предпочтения, причем за первое место проекту выставляется S баллов, за второе – S-1 и так далее (за последнее место — 1 балл). Кроме того, могут применяться следующие системы оценок:

- за первое место S-1 балл, за второе S-2, и т.д., за последнее – 0 баллов;
- за первое место 1 балл, за второе 1/2 балла, и т.д., за последнее – 1/S баллов.

После выставления оценок всеми экспертами все набранные проектами баллы суммируются

$$b_j = \sum_{i=1}^{M} r_{ij}\,, \qquad (9.4)$$

а победителем объявляется проект, набравший наивысший суммарный балл:

$$w = \max(b_j)\,. \qquad (9.5)$$

Несмотря на то, что метод Борда был разработан для выбора одного наилучшего проекта из множества, он позволяет определить места, занимаемые в голосовании остальными проектами в зависимости от количества набранных очков.

Согласно ***методу Кондорсе*** (*Condorcet*) победителем объявляется тот проект *x*, который побеждает любой другой проект при парном сравнении по правилу большинства, т.е. для всякого проекта $y \in S$, $y \neq x$, экспертов, ранжирующих эти проекты в порядке $x \succ y$ больше, чем тех, кто ранжирует их в порядке $y \succ x$.

Для каждого эксперта данные парные сравнения могут быть записаны в виде матрицы, пример которой приведен на рисунке 9.4. Здесь 1 обозначен факт предпочтения экспертом проекта-кандидата проекту-оппоненту, 0 – предпочтение проекта-оппонента проекту-кандидату, по главной диагонали ставятся прочерки, т.к. сравнивать проект сам с собой нецелесообразно.

		Оппонент			
		Проект 1	Проект 2	Проект 3	Проект 4
Кандидат	Проект 1	-	0	0	1
	Проект 2	1	-	1	1
	Проект 3	1	0	-	1
	Проект 4	0	0	0	-

Рисунок 9.4 – Пример матрицы сравнений одного эксперта при оценке проектов методом Кондорсе

После определения матриц сравнений для каждого из экспертов они складываются в общую матрицу сравнений, пример которой приведен на рисунке 9.5.

При попарном сравнении кандидата *i* и его оппонентов необходимо сравнивать значения *i*-й строки с соответствующими значениями *i*-го столбца:

$$c_{ij} > c_{ji}. \qquad (9.6)$$

Если для всех *j*=1..*M*, *j*≠*i* неравенство (9.6) выполняется, то проект *i* объявляется победителем.

		Оппонент			
		Проект 1	Проект 2	Проект 3	Проект 4
Кандидат	Проект 1	-	8	8	8
	Проект 2	13	-	10	21
	Проект 3	13	11	-	14
	Проект 4	13	0	7	-

Рисунок 9.5 – Пример общей матрицы сравнений при оценке проектов методом Кондорсе

Победитель по Кондорсе существует не всегда, т.к. при наличии более двух проектов и более двух экспертов коллективное ранжирование проектов

может быть цикличным (так называемый парадокс Кондорсе), то есть для каждого проекта-кандидата есть предпочитаемый большинством проект-оппонент.

С графической точки зрения выбор победителя по Кондорсе можно представить построением графа и приведением его к ярусно-параллельной форме. На этапе подготовки к построению графа результаты сравнения (9.6) можно записать в виде матрицы, для рассматриваемого примера представленной на рисунке 9.6.

		Оппонент			
		Проект 1	Проект 2	Проект 3	Проект 4
Кандидат	Проект 1	-	0	0	0
	Проект 2	1	-	0	1
	Проект 3	1	1	-	1
	Проект 4	1	0	0	-

Рисунок 9.6 – Матрица отношений проектов

В построенном по результатам сравнений графе стрелками указывается отношения между парами объектов, причем стрелка от i-го проекта к j-му указывает, что большинство экспертов отдали предпочтение i-му проекту по сравнению с j-м. Вид графа для рассматриваемого примера приведен на рисунке 9.7

Ярусно-параллельная форма представляет собой такую форму графа, где в каждом ярусе находятся вершины, в которые нет входящих стрелок из вершин, лежащих в нижних ярусах, а есть только входящие стрелки из вершин, находящихся на более высоких ярусах. При преобразовании рассматриваемого графа к ярусно-параллельной форме получим граф, представленный на рисунке 9.8.

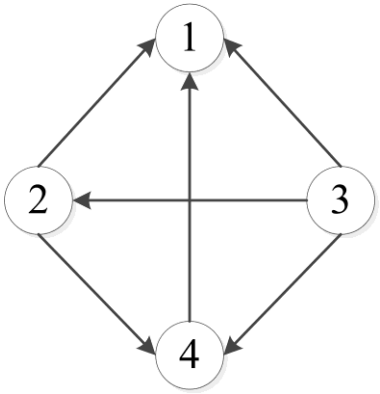


Рисунок 9.7 – Граф предпочтений проектов

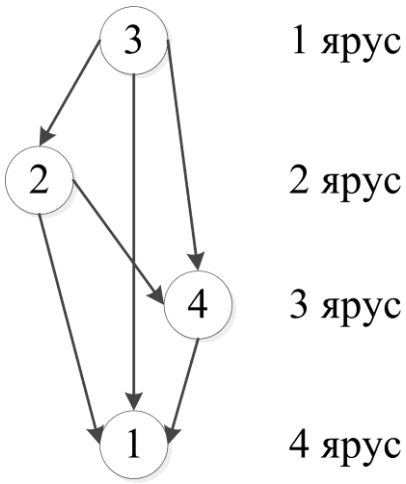


Рисунок 9.8 – Граф предпочтений проектов в ярусно-параллельной форме

Из анализа графа в ярусно-параллельной форме можно сделать вывод, что победителем является 3 проект, второе место занимает проект номер 2, третье – проект номер 4, и последнее, четвертое – 1 проект.

Вопросы к зачету

4. Виды экспериментальных исследований: наблюдение и эксперимент. Фактор и отклик. Виды факторов. Типы физических величин. Типы погрешностей измерений. Однократные и многократные измерения. Прямые и косвенные измерения. Оценка погрешности прямых многократных измерений. Оценка погрешности косвенных многократных измерений.

5. Случайная величина. Генеральная совокупность и выборка. Функция плотности вероятности. Интегральная функция вероятности. Принципы подбора выборки. Параметры распределения, влияющие на вид кривой распределения.

6. Нормальный закон распределения и его свойства. Равномерный закон распределения и его свойства. Логарифмически нормальный закон распределения и его свойства. Экспоненциальный закон распределения и его свойства. Распределение Вейбулла и его свойства.

7. Алгоритм построения гистограммы и полигона частот. Интерпретация вероятностных графиков. Алгоритм построения Q-Q графика. Алгоритм построения P-P графика.

8. Общий алгоритм проверки статистических гипотез. Алгоритм применения критерия согласия Колмогорова при проверке простой и сложной гипотезы на примере нормального закона распределения. Алгоритм применения критерия согласия Пирсона при проверке простой и сложной гипотезы на примере нормального закона распределения.

9. Алгоритм нахождения функции линейной одномерной регрессии. Коэффициент детерминации. Алгоритм проверки адекватности линейной одномерной регрессионной модели.

10. Коэффициент линейной корреляции. Корреляционное поле. Сила и направление корреляционной связи. Связь корреляционного и регрессионного анализов. Проверка гипотезы о значимости генерального коэффициента линейной корреляции. Проверка гипотезы о значимости различия между двумя коэффициентами линейной корреляции.

11. Основные понятия дисперсионного анализа. Предварительная оценка данных при дисперсионном анализе. Алгоритм применения критерия Кохрена о равенстве дисперсий. Алгоритм проведения однофакторного дисперсионного анализа. Алгоритм проведения двухфакторного дисперсионного анализа.

12. Методы голосования. Алгоритм метода "Медиана Кемени". Алгоритм метода Борда. Алгоритм метода Кондорсе.

ЛИТЕРАТУРА

1. Gibbons J. D., Chakraborti S. Nonparametric Statistical Inference, Fourth Edition: Revised and Expanded. Boca Raton, Florida: CRC Press, 2014. 680 p.

2. Thode H. C. Testing for normality. Boca Raton, Florida: CRC Press, 2002. 368 p.

3. Айвазян С.А., Енюков И.С., Мешалкин Л.Д. Прикладная статистика: основы моделирования и первичная обработка данных. Справочное издание. М.: Финансы и статистика, 1983. 471с.

4. Афифи А., Эйзен С. Статистический анализ: подход с использованием ЭВМ. М.: Мир 1982. 488 с.

5. Дрейпер Н., Смит Г. Прикладной регрессионный анализ: в 2-х кн. Кн.1. Изд. 2-е, перераб. и доп. М.: Финансы и статистика, 1986. 366 с.

6. Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников. М.: ФИЗМАТЛИТ, 2006. 816 с. ISBN 5-9221-0707-0.

7. Мухин О.И. Моделирование систем [Электронный ресурс] / Stratum. [сайт]. URL: http://stratum.ac.ru/education/textbooks/modelir/lection 36.html.

8. Орлов А.И. Оценивание для сгруппированных данных // Статистические методы оценивания и проверки гипотез: межвузовский сборник научных трудов. Пермский государственный университет. Пермь. 2012. Вып. 24. С. 83-95

9. Орлов А.И. Теория принятия решений. Учебное пособие. М.: Издательство "Март", 2004. 656 с.

10. Орлова А.А. Дисперсионный анализ факторов при имитационном моделировании сложных систем [Электронный источник] // Молодежный научно-технический вестник. 2013. № 11. URL: http://sntbul.bmstu.ru/doc/ 635566.html

11. Рекомендация МСЭ-R P.1057-2. Распределения вероятностей, касающихся моделирования распространения радиоволн / МСЭ: Верен идее соединить мир [сайт]. URL: http://www.itu.int/dms_pubrec/itu-r/rec/p/R-REC-P.1057-2-200708-S!!MSW-R.doc

12. Сотников Р. Использование методов экспертных оценок в оценочной практике [Электронный ресурс] // Центр экономики проектов. [сайт] URL: http://cepgroup.ru/backoffice/100-metod-expert-ozenka

13. Статистический анализ данных, моделирование и исследование вероятностных закономерностей. Компьютерный подход: монография / Б.Ю. Лемешко [и др.]. Новосибирск: Изд-во НГТУ, 2011. 888 с. ISBN 978-5-7782-1590-0.

14. Хабибуллин Р.Ф. Голосования и коллективный выбор: Учебное пособие. Казань: Казанский государственный университет, 2009. 28 с.

15. Харченко М.А. Корреляционный анализ. Воронеж: Издательско-полиграфический центр Воронежского государственного университета, 2008. 31 с.

16. Ходасевич Г.Б.. Обработка экспериментальных данных на ЭВМ. Часть I. Обработка одномерных данных [Электронный ресурс]: учебное пособие // URL: http://dvo.sut.ru/libr/opds/i130hodo_part1/index.htm.

17. Хомич А.В. Элементарные основы математической статистики и компьютерная обработка данных в психологии [Электронный ресурс]: учебно-методическое пособие для очно-заочного и заочного отделения / URL: http://khomich.narod.ru/metodichka/Kriterii/Kriterii.htm