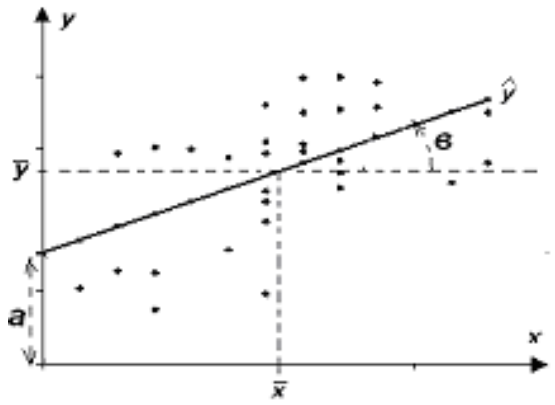
Л.В. Филатов

ЗАДАЧИ СТАТИСТИЧЕСКОГО АНАЛИЗА В СТРОИТЕЛЬСТВЕ

Корреляционный, регрессионный и факторный анализ

Учебно-методическое пособие



Нижний Новгород
2017

Министерство образования и науки Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Нижегородский государственный архитектурно-строительный университет»

Л.В. Филатов

ЗАДАЧИ СТАТИСТИЧЕСКОГО АНАЛИЗА В СТРОИТЕЛЬСТВЕ

Корреляционный, регрессионный и факторный анализ

Утверждено редакционно-издательским советом университета
в качестве учебно-методического пособия

Нижний Новгород
ННГАСУ
2017

ББК 22.172
Ф 51
УДК 530.1

Рецензенты:

Д.Н. Шуваев – к.т.н., доцент кафедры ТиМДО ННГУ им. Н.И. Лобачевского
С.Н. Охулков – к.ф-м.н., доцент кафедры ТОЭ НГТУ им. Р.Е. Алексеева

Филатов Л.В. Задачи статистического анализа в строительстве. Корреляционный, регрессионный и факторный анализ [Текст]: учеб.-метод. пособие /Л.В. Филатов; Нижегор. гос. архитектур. - строит. ун - т – Н. Новгород: ННГАСУ, 2017. – 68 с.
ISBN 978-5-528-00223-1

Рассматривается задачи статистического анализа, возникающие в различных областях строительства, таких как анализ характеристик строительных материалов, проектирование, возведение и эксплуатация строительных конструкций, экономике строительного производства, маркетинговых исследованиях и ряда других областей. Статистический анализ это подход к исследованию объектов различной природы с множеством функциональных и структурных связей, усложненных различными неопределенностями и рисками. В пособии рассматривается теоретический материал, снабженный множеством примеров.

Может быть использовано для подготовки к лекционным и практическим занятиям, а также для самостоятельного выполнения расчетной работы, варианты которых предлагаются в конце.

Предназначено для обучающихся по дисциплинам «Теория вероятностей и математическая статистика», «Обработка результатов измерений», «Прикладные задачи математики в строительстве», «Факторный анализ» и другие.

ISBN 978-5-528-00223-1

Введение

Задачи статистического анализа очень часто встречаются в различных областях научно-практической деятельности человека. Эти задачи связаны со сбором и обработкой наблюдательных данных над теми или иными объектами, явлениями или процессами. Под наблюдениями понимается довольно широкое понятие, включающее в себя получение наборов статистических данных различного свойства и форматов. Объекты таких наблюдений могут иметь различную природу, это, например, экспериментальные измерения в научно-технических установках, эксплуатационно-технические проверки и акты, геологические паспорта территорий, финансово-экономические отчеты, регистрационно-кадастровые документы, социально-политические опросы, медико-психологические обследования и многое другое. Обработка статистических данных связана с анализом и выявлением закономерностей в поведении объектов с целью объяснений наблюдений, выявления причин и предсказания поведения наблюдаемых объектов.

В области строительства также имеется ряд задач по обработке и анализу о свойствах строительных материалов и конструкций, отраженных в текущих и аттестационных поверочных измерениях, отчетах о себестоимостях строительства объекта и его содержания, о состоянии инфраструктурных элементов.

В связи со столь широким предметом исследования статистика превратилась из наблюдательно-фиксирующей науки в аналитическую с широчайшим применением математических методов планирования, сбора и обработки статистических данных. Наблюдаемые величины всегда в той или иной степени являются непредсказуемыми, то есть случайными. Случайность измеряемых величин связана как с их внутренней стохастичностью, так и с ошибками измерения, вносимыми измеряющим прибором и субъектом. Методы теории вероятностей и математической статистики лежат в основе всех методов статистического анализа.

К задачам статистического анализа [1-3] обычно относят следующие задачи:

- корреляционный анализ (определение зависимости величин),
- регрессионный анализ (определение формы зависимости),
- дисперсионный анализ (анализ влияния условий измерения),
- факторный анализ (определение наиболее значимых факторов),
- кластерный анализ (классификация и идентификация объектов),
- дискриминантный анализ (выбор наилучших решений),
- анализ временных рядов (обработка изменяющихся во времени данных, их сглаживание и прогноз),
- планирование эксперимента (анализ мероприятий, необходимых для достижения максимальной точности измерений и достоверности выводов),
- фрактальный анализ (выделение внутренних структур в объектах).

В настоящем пособии, предназначенном для студентов различных специальностей, излагаются методы решения задач корреляционного, регресси-

онного и факторного анализа. Для решения задач статистического анализа имеется ряд пакетов прикладных программ, в данном пособии опираемся на статистический пакет универсального приложения Excel-13.

В главах 1 и 2 кратко рассматриваются основные понятия и методы теории вероятностей и математической статистики. Наибольшее внимание уделено применению основных статистических методов, таких как выборочный метод, методы статистических оценок, методы проверки статистических гипотез.

Глава 3 посвящена многомерным статистическим данным. В ней рассматриваются типы и форматы наблюдательных данных, их представления и преобразования. Анализируется их засоренность грубыми ошибками измерений.

Вопросы взаимосвязи между наблюдаемыми величинами рассмотрены в главе 4, устанавливается наличие значимой корреляционной связи между измеренными величинами.

Вид этих связей рассматривается в главе 5, где методами регрессионного анализа наблюдаемых данных с использованием метода наименьших квадратов, устанавливается значимая линейная или нелинейная связь переменных. Качество регрессионной зависимости анализируется на предмет выполнения предпосылок метода наименьших квадратов, сформулированных в теореме Гаусс-Маркова.

Глава 6 посвящена поиску новых скрытых от непосредственного измерения переменных, имеющих большое значение для анализа наблюдаемых данных. Методами факторного анализа проводится факторизация модели главных координат измеряемых величин, определяются главные факторы, несущие на себе подавляющую долю изменчивости наблюдаемых переменных.

На протяжении всех глав в качестве примера рассматривается задача статистического анализа многомерного статистического набора измеренных при наблюдении величин. Решение ведется при помощи статистического пакета Excel.

В пособии даются варианты выполнения самостоятельных практических работ для студентов по решению задач анализа многомерных статистических наборов данных.

В приложении приводятся справочные данные по распределению случайных величин, критериям проверки значимостей, и др.

1. Случайные величины

Теория вероятностей - математическая наука о случайных явлениях окружающего нас мира, имеющая серьезное эмпирическое обоснование. Случайность, как неоднозначность и непредсказуемость явлений обусловлена их сложностью или их квантовой сущностью, а также субъективным действием или восприятием человека. Рассмотрим ***Опыт*** как любое наблюдение (созерцание, измерение или эксперимент) случайного явления в произвольной практической деятельности человека, например, в бытовой, научно-познавательной, производственно-технической, социально-экономической, психофизической или какой-либо другой сфере деятельности. ***Событием*** будем называть любой возможный исход опыта. Множество всех событий опыта образуют его ***событийное пространство***. События в опыте могут быть как равносильными, так и неравносильными с точки зрения их наступления в опыте. Числовая величина, характеризующая возможность наступления события в опыте называется ***вероятностью*** события. Определение (вычисление) вероятности события производится обычно через его ***частоту*** наблюдения при массовом повторении опыта, без изменения условий проведения. Иногда, в случае наличия симметрии опыта, возможно определение теоретической (до опытной) вероятности, через число ***равновозможных*** исходов опыта, приводящих к наступлению события. Например, пусть А событие, состоящее в выпадении «шестерки» в опыте по бросанию игровой кости, тогда *P(A)=1/6* есть **теоретическая** вероятность события при бросании правильной кости (симметричной без смещения центра тяжести). Если кость неправильная (изношенная, специально изготовленная), то можем определить только частоты события $\nu_{36}(A) = 5/36$, $\nu_{72}(A) = 9/72$, .., при 36, 72, .. однотипных бросаниях такой кости. Но если отклонения этих частот с ростом числа бросаний уменьшаются, то это означает, что мы с определенной точностью вычислим ***эмпирическую*** **(**статистическую**)** вероятность события.

Другим важнейшим понятием теории вероятностей, после понятия события, является понятие ***случайной величины***.

1.1. Понятие и описание случайных величин

Внимательно анализируя опыт, можно заметить, что помимо событий в опыте обычно можно увидеть и ввести некоторую числовую величину, которая своими значениями описывает все множество событий опыта. Например:

I - число, выпадающее на игральной кости. $I \in \{1,2,3,4,5,6\}$

N - число посетителей сайта за сутки. $N \in \{1,2,3,...\}$

T_S - время работы устройства до первого сбоя. $T_S \in \{[0;+\infty)\}$

Случайной величиной называется числовая величина, принимающая в опыте случайным образом одно и только одно значение из всех своих возможных значений. Будем обозначать случайные величины большими латинскими буквами X, Y, Z,..., а их возможные значения малыми $x, y, z, \ldots$.

Множество всех возможных значений случайной величины X будем обозначать $\Omega_X = \{x\}$, в зависимости от вида этого множества случайные величины делятся на ***дискретные*** и ***непрерывные***. Дискретная величина принимает конечное или бесконечное, но счетное, число значений

$$\Omega_X = \{x_1, x_2, ..., x_n, ...\},$$

а непрерывная величина принимает значения из конечного или бесконечного непрерывного числового интервала

$$\Omega_X = \{(a, b)\} \quad -\infty \le a < b \le +\infty.$$

Случайная величина проявляется в опыте через свои значения, поскольку каждое значение есть события $A = (X = x)$, $B = (X > x)$ и др. В связи с этим необходимо уметь вычислять вероятности этих событий, то есть вероятность того, что случайная величина принимает те или иные значения.

Законом распределения $P_X(x)$ случайной величины называют любое правило (функция, таблица, график, алгоритм,..), которое устанавливает соответствие между возможными значениями случайной величины и вероятностями, с которыми она принимает эти значения.

Задание случайной величины X, области ее возможных значений Ω_X и закона распределения $P_X(x)$ полностью определяют случайную величину как ***вероятностную модель*** случайного явления, наблюдаемого в опыте.

Поскольку у случайной дискретной величины все значения можно перечислить, то ее закон распределения удобно задавать в виде таблицы вероятностей для упорядоченных значений величины:

Значения случайной величины (X)	x_1	x_2	x_3	…	x_n	…
Вероятности значений (P)	p_1	p_2	p_3	…	p_n	…

Получается так называемый ***ряд распределения*** случайной дискретной величины. Причём для вероятностей всех событий $p_k = P(X = x_k)$ выполнено:

$$p_1 + p_2 + p_3 + \ldots + p_n = \sum_k p_k = 1,$$

что является *необходимым условием для закона распределения*.

Закон распределения случайной дискретной величины может быть задан и ***функционально*** в виде $p_m = p(x_m)$, а часто в графическом виде в форме так называемого ***многоугольника распределения вероятностей*** случайной дискретной величины, изображенного на рис. 1.1.

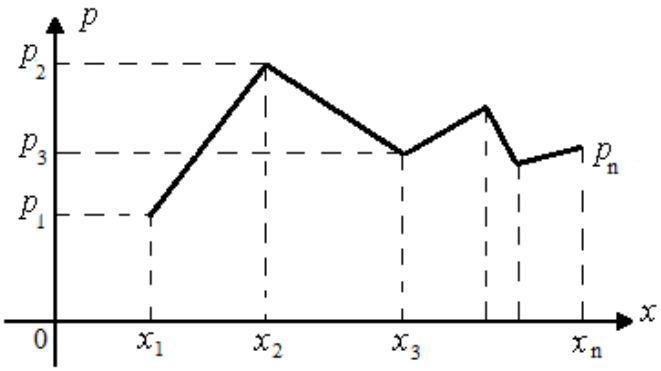


Рис. 1.1. Многоугольник распределения дискретной величины

Для случайной непрерывной величины невозможно говорить о вероятности значения случайной величины в точке $P(X = x)$, но можно определить вероятность ее значения в любом интервале области возможных значений

$$\Omega_X = \{(a,b)\} \quad -\infty \le a < b \le +\infty.$$

Функцией распределения случайной величины X называется функция $F_X(x)$, выражающая для каждого числа x из области возможных значений вероятность того, что случайная величина X примет значение, меньшее этого числа:

$$F(x) = P(X \le x), \ \forall x \in \Omega_X.$$

Функция распределения $F(x)$ принимает значения на отрезке $[0;1]$, т.к. ее значения есть вероятность события. Она будет рассматриваться как непрерывная и дифференцируемая функция, обладающая следующими важными свойствами [4,5]:

$P(x_1 \le X \le x_2) = F(x_2) - F(x_1)$,

$F(x_2) - F(x_1) \ge 0$, т.е. $F(x_2) \ge F(x_1)$ при $x_2 > x_1$.

$F(\tilde{o}) \xrightarrow[\tilde{o} \to a+0]{} 0, \quad F(x) \xrightarrow[x \to b-0]{} 1.$

Таким образом, функция распределения $F(x)$ не убывает, её значения расположены на отрезке $[0;1]$. При стремлении $x \to a$ функция распределения обращается в ноль, а при стремлении $x \to b$ функция распределения обращается в единицу. Примерный график функции распределения $F(x)$ приведён на рис. 1.2.

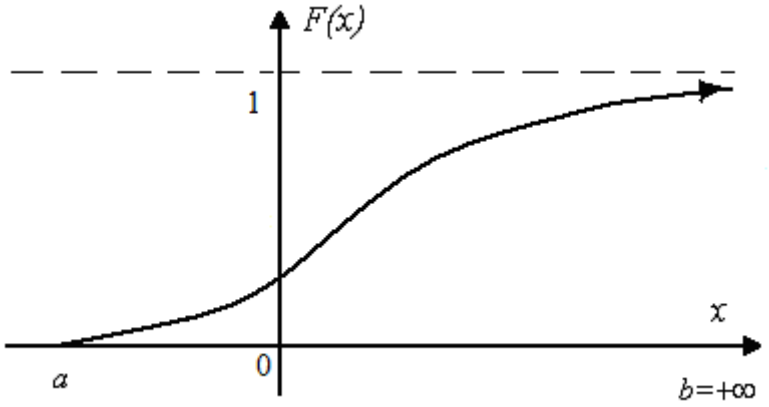


Рис. 1.2. Функция распределения случайной непрерывной величины

Пусть имеется непрерывная случайная величина, определённая в области $x \in \Omega_X = \{(a,b)\}\ -\infty < a < b < +\infty$ и описывается непрерывной и дифференцируемой функцией распределения $F(x)$. Вычислим вероятность нахождения случайной величины в h-интервале и поделим ее на длину интервала:

$$\frac{P_h(x \le X \le x+h)}{h} = \frac{F(x+h)-F(h)}{h} \xrightarrow[h\to 0]{} F'(x).$$

Такие величины называются обычно погонной плотностью или просто плотностью величины. ***Плотностью распределения вероятностей*** *(*или сокращённо ***плотностью вероятности)*** непрерывной случайной величины называется производная от её функции распределения: $f(x) = F'(x)$.

Плотность вероятности обладает рядом замечательных свойств [4,5]:

$$f(x) \ge 0, \quad \int_a^b f(x)\,dx = 1\ , \quad f(x) \xrightarrow[x\to a,b]{} 0,$$

$$P(x_1 \le X \le x_2) = \int_{x_1}^{x_2} f(x)\,dx, \quad F(x) = \int_a^x f(t)\,dt\ .$$

В силу указанных свойств, функция $f(x)$ плотности распределения вероятностей всегда неотрицательна, стремится к нулю на границах области возможных значений, вероятность нахождения в интервале значений величины равна площади под графиком функции $f(x)$, опирающейся на интервал значений, а вся площадь между графиком функции $f(x)$ и осью абсцисс равна единице. Примерный график функции $f(x)$ плотности распределения вероятностей изображён на следующем рис. 1.3.

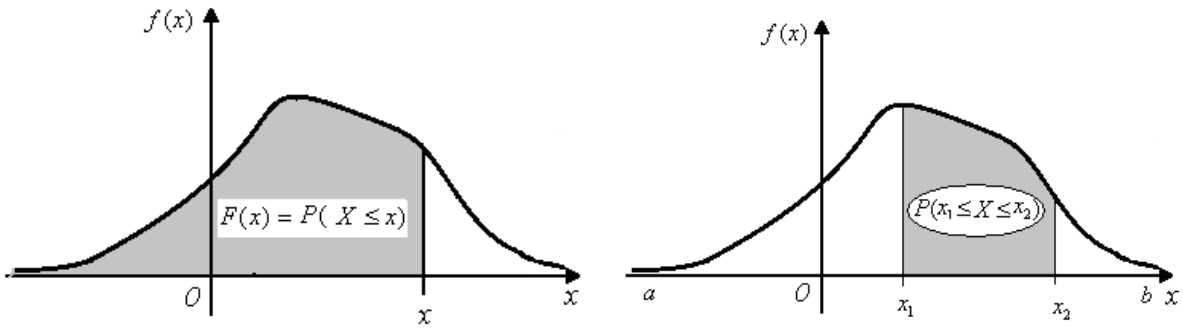


Рис. 1.3. Свойства функции распределения и плотности распределения

Итак, для полной характеристики случайной величины достаточно знать или функцию распределения, или плотность распределения вероятностей (т.к. одну из них можно выразить через другую):

$$F(x) = \int_a^x f(t)\,dt \quad \text{или} \quad f(x) = F'(x).$$

Часто, особенно в задачах математической статистики, удобнее использовать не функцию распределения $F(x)$, а обратную к ней функцию $F_{обр}(p)$, которая, как и сама функция распределения является монотонной, однозначной и непрерывной функцией от вероятности. Так для выделения части области возможных значений Ω_X, где случайная величина может находиться

(принимать эти значения в опыте) с той или иной заданной вероятностью, используются квантили распределения по заданному уровню вероятности.

Левосторонняя квантиль $x_\beta = F_{îáð}(\beta)$ определяется как $P(X \le x_\beta) = \int_a^{x_\beta} f(x)\,dx = \beta$,

а правосторонняя квантиль $x_\alpha = F_{îáð}(1-\alpha)$ определяется $P(X \ge x_\alpha) = \int_{x_\alpha}^{b} f(x)\,dx = \alpha$.

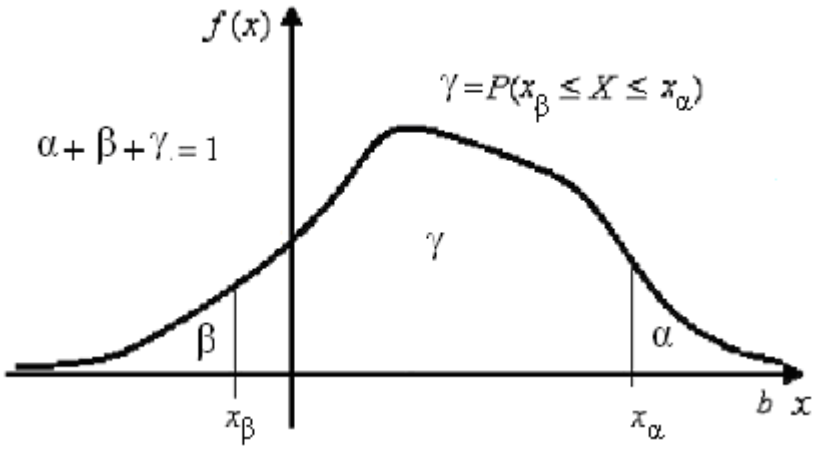


Рис. 1.4. Левосторонняя и правосторонняя квантили распределения

Обе эти квантили используются для отсечения приграничных частей у области возможных значений $\Omega_X = \{(-\infty,+\infty)\}$ рис. 1.4, а для выделения срединной части области часто используется центральная квантиль [6], где случайная величина будет находиться с вероятностью γ. Границы центральной квантили, за которые случайная величина выходит с равной вероятностью $\alpha = \beta = (1-\gamma)/2$:

$$x_{\gamma 1} = F_{îáð}\left(\frac{1-\gamma}{2}\right),\ x_{\gamma 2} = F_{îáð}\left(\frac{1+\gamma}{2}\right),\ P(x_{\gamma 1} \le X \le x_{\gamma 2}) = \int_{x_{\gamma 1}}^{x_{\gamma 2}} f(x)\,dx = \gamma.$$

Пример. Рассмотрим случайную непрерывную величину, определённую на конечном отрезке с линейно нарастающей функцией распределения рис. 1.5.

$$F(x) = \frac{x-a}{b-a},\quad f(x) = \frac{1}{b-a}\quad ,\quad x \in \Omega_X = \{(a,b)\}\quad -\infty < a < b < +\infty$$

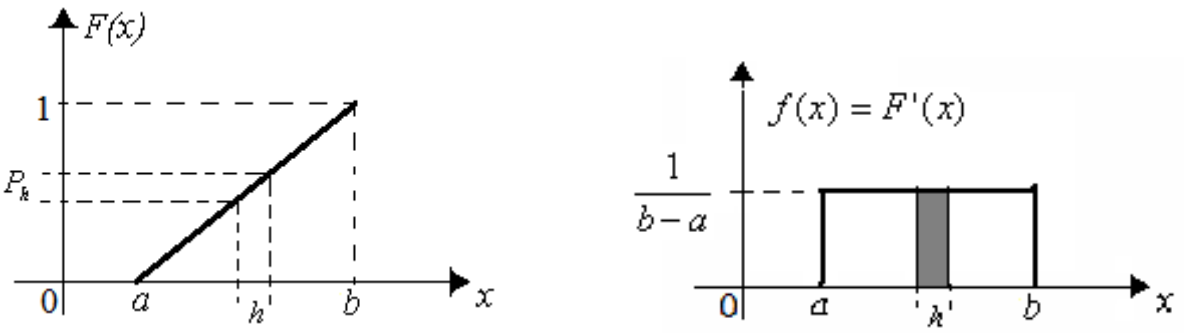


Рис. 1.5. Функция распределения равномерной случайной величины

Вычисляя вероятность попадания случайной величины в интервал длиной h, получим, что $P_h = P(x < X < x+h) = F(x+h) - F(x) = h/(b-a)$, то есть эта вероятность постоянна и не зависит от расположения h интервала в области Ω_X, а зависит лишь от длины интервала. Такая величина, принимающая свои значения с равной вероятностью во всех точках области возможных значений, называется ***равномерной случайной величиной***. Обратная функция распределения и квантили для такого распределения следующие:

$$F_{\text{îáð}}(p) = a + p(b-a),\ x_\beta = F_{\text{îáð}}(\beta),\ x_\alpha = F_{\text{îáð}}(1-\alpha).$$

1.2. Числовые характеристики случайных величин

Закон распределения случайной величины, заданный в той или иной форме, полностью определяет случайную величину как некоторую модель наблюдаемого в опыте явления. Однако часто в практической деятельности знание закона бывает невозможным, а то и избыточным, достаточно знать лишь некоторые общие (интегральные) характеристики случайной величины.

Пусть случайная величина X, дискретная или непрерывная, задается законом распределения, тогда основными характеристиками случайной величины являются:

Математическое ожидание:

$$M(X) = m_X = \begin{cases} \sum\limits_k x_k \cdot p_k & -\textit{для дискретной случайной величины} \\ \int\limits_{\Omega_x} x f_X(x)dx & -\textit{для непрерывной случайной величины} \end{cases}$$

Дисперсия:

$$D(X) = D_X = \begin{cases} \sum\limits_m (x_k - m_X)^2 \cdot p_k & -\textit{для дискретной случайной величины} \\ \int\limits_{\Omega_x} (x - m_X)^2 \cdot f_X(x)dx & -\textit{для непрерывной случайной величины} \end{cases}$$

Среднеквадратическое отклонение:

$$\sigma(X) = \sigma_X = \sqrt{D_X}.$$

Математическое ожидание $M(X) = m_X$ характеризует центр распределения или средневзвешенное ожидаемое значение величины, а геометрически оно изображается как координата центра тяжести фигуры, образованной осью x и линией функции $f(x)$ или $p(x_m)$. Дисперсия $D(X) = \sigma_x^2$ характеризует средний ожидаемый разброс (широту, изменчивость, вариативность) значений величины возле $M(X)$, поскольку совпадает с математическим ожиданием квадрата отклонения случайной величины от его математического ожидания.

$$D(X) = M(\tilde{X}^2), \text{ где } \tilde{X} = X - m_X.$$

Среднеквадратическое отклонение $\sigma(X) = \sigma_x$ имеет тот же смысл, что и дисперсия, но в отличие от неё имеет размерность, совпадающую с размерно-

стью самой случайной величины, что более удобно и позволяет изобразить его как и математическое ожидание на рис. 1.6. Между дисперсией и математическим ожиданием имеется простая связь $D(X) = M(X^2) - M^2(X)$.

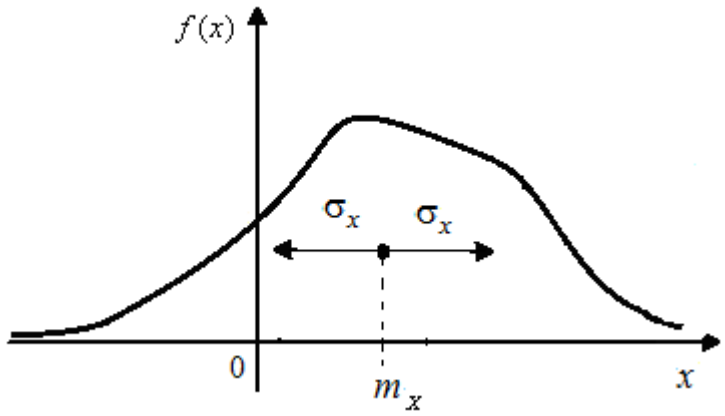


Рис. 1.6. Геометрическая иллюстрация понятий математического ожидания $M(X) = m_X$ и дисперсии $D(X) = \sigma_x^2$ случайной величины

Пример. Рассмотрим случайную величину X, определенную на множестве возможных значений $\Omega_X = \{[0;+\infty)\}$ со следующим законом распределения $F_X(x) = 1 - e^{-\lambda x}$, $f_X(x) = \lambda \cdot e^{-\lambda x}$, где параметр $\lambda > 0$. Такая случайная непрерывная величина называется ***показательной*** **(**рис. 1.7).

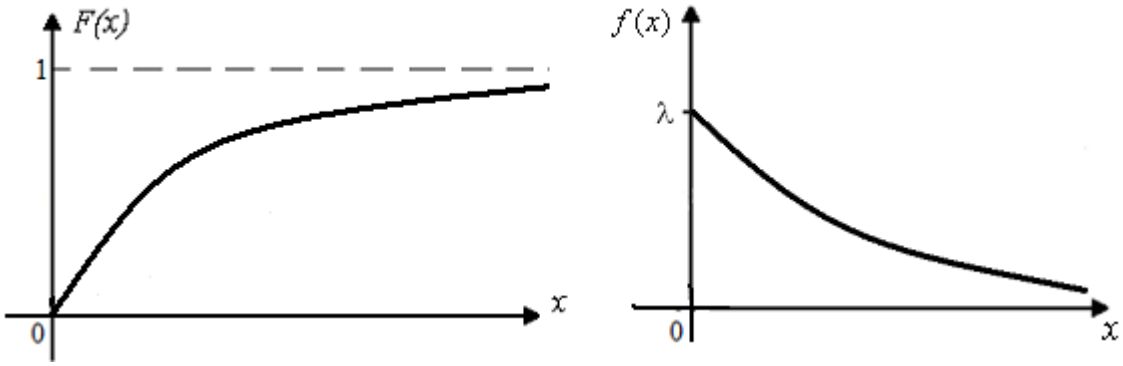


Рис. 1.7. Функция распределения $F_X(x)$ и плотность распределения $f_X(x)$ показательной случайной величины

$$M(X) = \int_0^\infty x \cdot \lambda e^{-\lambda x} dx = -x \cdot e^{-\lambda x}\Big|_0^\infty + \int_0^\infty e^{-\lambda x} dx = 0 - \frac{1}{\lambda} e^{-\lambda x}\Big|_0^\infty = \frac{1}{\lambda}$$

$$M(X^2) = \int_0^\infty x^2 \cdot \lambda e^{-\lambda x} dx = -x^2 \cdot e^{-\lambda x}\Big|_0^\infty + \frac{2}{\lambda} \cdot \int_0^\infty x \cdot \lambda e^{-\lambda x} dx = 0 - \frac{2}{\lambda} \cdot \frac{1}{\lambda} = \frac{2}{\lambda^2}$$

$$D(X) = M(X^2) - M^2(X) = \frac{2}{\lambda^2} - \left(\frac{1}{\lambda}\right)^2 = \frac{1}{\lambda^2}$$

Математическое ожидание постоянной величины равно этой постоянной величине, а её дисперсия равна нулю:

$$X = C = const \Rightarrow M(X) = C,\ D(X) = 0.$$

Умножение случайной величины на постоянный множитель приводит к следующему изменению её характеристик:

$$M(C\cdot X)=C\cdot M(X),\ D(C\cdot X)=C^2\cdot D(X),\ \text{где } C=const.$$

Математическое ожидание суммы конечного числа случайных величин равно сумме математических ожиданий этих величин:

$$M(X_1+X_2+\ldots+X_k)=M(X_1)+M(X_2)+\ldots+M(X_k).$$

Из вышеприведённых свойств можно заметить, что при преобразовании случайной величины X по линейному закону в величину Y

$$Y=\frac{X-m_X}{\sigma_X}\ \Rightarrow\ M(Y)=0,\ D(Y)=1.$$

Такое преобразование случайной величины называется центрированием и нормированием, а характеристики получаемой величины называются стандартными.

Для независимых случайных величин X и Y имеет место:

$$D(X+Y)=D(X)+D(Y),$$
$$M(XY)=M(X)\cdot M(Y).$$

Величины называются ***независимыми***, если распределение любой из них не зависит от того, какие значения принимает другая величина. В противном случае величины являются статистически зависимыми.

1.3. Нормальная случайная величина

Случайная величина имеет ***нормальный закон распределения*** (закон **Гаусса**), если она определена в области $\Omega_X=\{(-\infty;+\infty)\}$, а её плотность распределения вероятностей имеет вид:

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-m)^2}{2\sigma^2}}=\text{Í Î ÐÌ .ÐÀÑÏ}\ (x,m,\sigma,0)$$

где m и σ - параметры распределения $(\sigma>0,-\infty<m<+\infty)$.

Нормальный закон распределения $X=N(m,\sigma)$ наиболее часто встречается на практике. Главная его особенность – он является предельным законом, которым приближаются другие, более сложные законы распределения [7].

Плотность вероятности $f(x)$ похожа на «колокол» (рис. 1.8).

При уменьшении только параметра σ, график функции сжимается и поднимается вверх по оси ординат. При изменении только параметра m, график перемещается вдоль оси абсцисс.

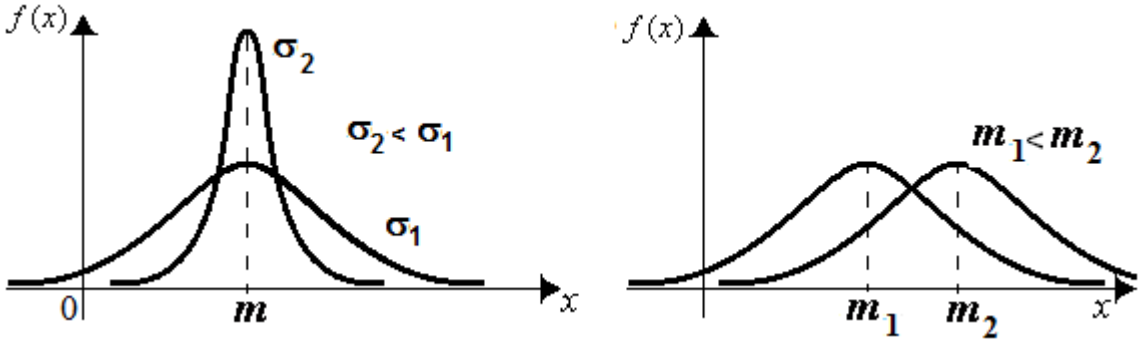


Рис. 1.8. Функция плотности распределения нормальной величины

Функция распределения $F(x)$ нормальной величины имеет вид:

$$F(x)=\int_{-\infty}^{x} f(t)\,dt=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x} e^{-\frac{(t-m)^2}{2\sigma^2}}\,dt=\frac{1}{2}+\hat{O}\left(\frac{x-m}{\sigma}\right)=\text{Í Î ÐÌ .ÐÀÑÏ }(x,m,\sigma,1),$$

где $\Phi(x)=\frac{1}{\sqrt{2\pi}}\int_{0}^{x} e^{-\frac{u^2}{2}}\,du$ - функция Лапласа. График функции распределения $F(x)$ изображен на рис. 1.9.

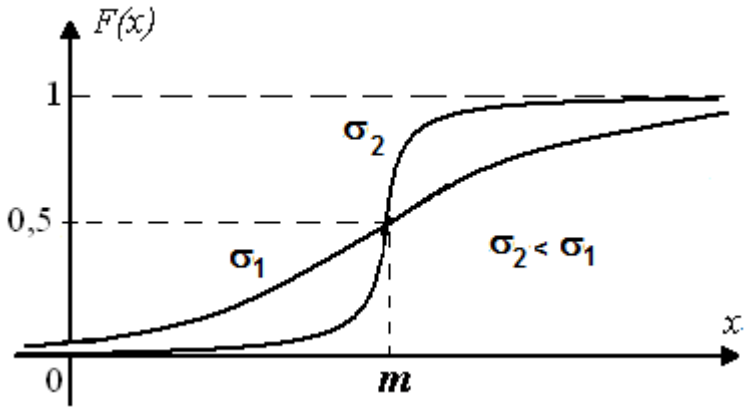


Рис. 1.9. Функция распределения нормальной величины

Вероятность того, что изучаемая случайная величина (распределённая нормально) примет значение в пределах от x_1 до x_2 вычисляется по обычной формуле:

$$P(x_1\le X\le x_2)=\hat{O}\left(\frac{x-m}{\sigma}\right)\Big|_{x_1}^{x_2}=\hat{O}\left(\frac{x_2-m}{\sigma}\right)-\hat{O}\left(\frac{x_1-m}{\sigma}\right).$$

В частном случае, когда интервал симметричен относительно точки m, эта формула выглядит так:

$$P(m-\varepsilon\le X\le m+\varepsilon)=\ P(|X-m|\le\varepsilon)=2\Phi\left(\frac{\varepsilon}{\sigma}\right).$$

Рассмотрим вероятность того, что изучаемая случайная величина (распределённая нормально) примет значение в пределах от $m-3\sigma$ до $m+3\sigma$:

$$P(m-3\sigma\le X\le m+3\sigma)=2\Phi(3\sigma/\sigma)=2\Phi(3)\approx 0{,}9973,$$

т.е. вероятность значений изучаемой случайной величины именно на интервале $[m-3\sigma,\,m+3\sigma]$ велика. Это утверждение составляет правило «трёх сигм». Числовые характеристики нормальной случайной величины будут:

$$M(X)=m,\ D(X)=\sigma^2.$$

Пример. Наблюдение за скоростью автомашин на определённом участке дороги показало, что скорость есть нормальная случайная величина с математическим ожиданием 60 км/ч и среднеквадратическим отклонением 10 км/ч. Определить вероятность того, что:

- скорость на этом участке не превышает 80 км/ч,

- скорость не отклоняется от математического ожидания более чем на 20%.

Поскольку скорость есть нормальная величина с параметрами $m=60$ и $\sigma=10$, то по основным формулам находим:

$$P(0 \le V \le 80) = \hat{O}(\frac{x-m}{\sigma})\Big|_0^{80} = \hat{O}(\frac{80-60}{10}) - \hat{O}(\frac{0-60}{10}) = \hat{O}(2) - \hat{O}(-6) = 0,477 + 0,5 = 0,947,$$

$$20\%\ \ \hat{i}\,\grave{o}\quad 60 = 12 \qquad P(|V-60| \le 12) = 2\Phi(\frac{12}{10}) = 2\hat{O}(1,2) = 2 \cdot 0,385 = 0,770.$$

Вычислим скорость, которую автомашины на этом участке не превышают с вероятностью 0,99. Из уравнения

$$P(0 \le V \le v_{\max}) = \hat{O}(\frac{x-m}{\sigma})\Big|_0^{v_{\max}} = \hat{O}(\frac{v_{\max}-60}{10}) - \hat{O}(\frac{0-60}{10}) = \hat{O}(\frac{v_{\max}-60}{10}) - \hat{O}(-6) = 0,99,$$

$$\hat{O}(\frac{v_{\max}-60}{10}) = 0,49 \ \Rightarrow\ \frac{v_{\max}-60}{10} = 2,33\ v_{\max} = 60 + 10 \cdot 2,33 = 83,3.$$

1.4. Системы случайных величин

Если рассматривается система случайных величин $X, Y, Z, ..$, то между ними могут быть следующие взаимные соотношения:

- они могут быть независимыми, когда распределение каждой из них не зависит от того, какие значения примут другие величины. Например, X - температура воды на входе системы отопления жилого многоквартирного дома, а Y - количество жильцов, проживающих в доме, эти величины независимы;

- они могут быть зависимы функционально, когда между значениями величин имеется функциональная связь вида $Y=\varphi(X)$. Так, площадь выражается через измерения случайных размеров. Связь между распределениями величин устанавливается достаточно просто при взаимно однозначной функциональной связи [4]:

$$F_Y(y) = F_X(\psi(y)), \quad f_Y(y) = \psi'(y) f_X(\psi(y)),$$

где $\psi(y)$ обратная для $\varphi(x)$ функция. Например, для равномерной X и $Y = X^2$:

$$X = Rn(a,b),\ \ b > a > 0,\ \psi(y) = \sqrt{y},\ F_Y(y) = \frac{\sqrt{y}-a}{b-a},\ f_Y(y) = \frac{1}{2\sqrt{y}} \cdot \frac{1}{b-a};$$

- случайные величины могут быть зависимыми статистически, когда распределение каждой случайной величины зависит от того, какие значения принимают другие величины. Например, X -температура воды на входе системы отопления жилого многоквартирного дома, а Y - количество жильцов, обратившихся с жалобой в ДУК на холод в квартирах , эти величины зависимы статистически.

Такая зависимость полностью может быть описана условными распределениями величин. Так, для пары величин X, Y условное распределение задаётся функцией двух переменных $f_X(x|y)$ или $f_Y(y|x)$, представляющих собой распределения одной величины при заданном значении другой величины. Распределения самих величин связаны с условными распределениями следующим образом:

$$f_X(x) = \int_{\Omega_Y} f_X(x|y)dy, \quad f_Y(y) = \int_{\Omega_X} f_Y(y|x)dx,$$

причем, оказывается, что $f_X(x)f_Y(y|x) = f_Y(y)f_X(x|y) = f(x,y)$, а $f(x,y)$ называется функцией совместного распределения и она связана с вероятностью значений величин через функцию совместного распределения

$$F(x,y) = P(X \le x, Y \le y) = \int\limits_{X \le x, Y \le y} f(x,y)dxdy .$$

Часто рассматриваются условные математические ожидания величин

$$\bar{Y}(x) = M_Y(x) = \int\limits_{\Omega_Y} y \cdot f_y(y|x)dy , \quad \bar{X}(y) = M_X(y) = \int\limits_{\Omega_x} x \cdot f_x(x|y)dx ,$$

такая зависимость средних значений (математических ожиданий) от значения других переменных называется регрессией. Функция регрессии $g(x) = \bar{Y}(x)$ и условные распределения иллюстрируются на рисунке 1.10.

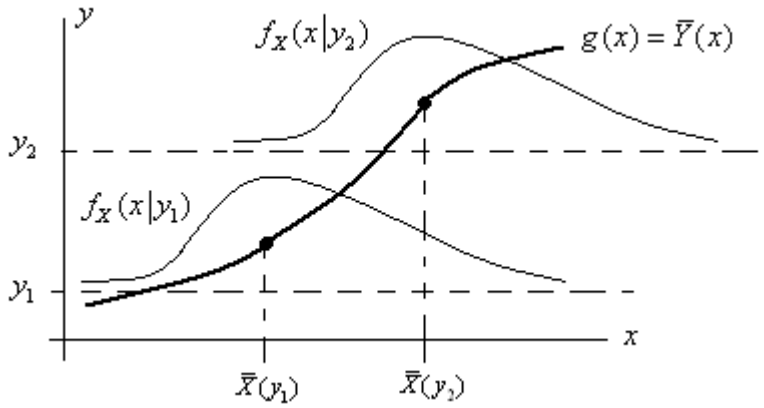


Рис. 1.10. Функция регрессии для зависимых величин

В случае независимости величин условные распределения совпадают и $f_Y(y|x) = f_Y(y)$, $f_X(x|y)f_X(x)$, а $f(x,y) = f_X(x) \cdot f_Y(y)$, $M(x \cdot y) = M(x) \cdot M(y)$.

В случае статистической зависимости введём понятие ковариационного момента (ковариации):

$$Cov(X,Y) = M(X \cdot Y) - M(X) \cdot M(Y),$$

который показывает степень статистической зависимости величин X и Y, поскольку при независимости переменных он равен нулю, а для статистически зависимых величин справедливы следующие формулы:

$$M(X \cdot Y) = M(X) \cdot M(Y) + Cov(X,Y), \quad D(X+Y) = D(X) + D(Y) + 2Cov(X,Y).$$

Введем также безразмерную величину коэффициента корреляции

$$\rho_{XY} = \frac{Cov(X,Y)}{\sqrt{D(X) \cdot D(Y)}},$$

обладающего следующими свойствами:

- его значение по модулю не превышает единицы $-1 \le \rho_{XY} \le 1$.
- для независимых величин X и Y $\rho_{XY} = 0$,
- для линейно зависимых величин $\rho_{XY} = \pm 1$.

Это позволяет использовать коэффициент корреляции в качестве меры статистической зависимости случайных величин. Говорят, что величины коррелируют между собой, если коэффициент корреляции не равен нулю.

2. Основные задачи и методы математической статистики

Для установления закономерностей, которым подчинены случайные события и случайные величины, теория вероятности, как и любая другая наука, обращается к опыту – наблюдениям, измерениям, экспериментам. Результаты наблюдений за случайными величинами объединяются в наборы статистических данных. Задачей математической статистики, раздела современной теории вероятностей, является разработка методов сбора и обработки статистических данных, а также их анализа с целью установления законов распределения наблюдаемых случайных величин [2].

2.1. Выборочный метод

Генеральной совокупностью является набор всех мыслимых статистических данных, при наблюдениях случайной величины:

$$x_{\Gamma} = \{x_1, x_2, x_3, ..., x_N\} = \{x_i; i = 1, N\}.$$

Наблюдаемая случайная величина X называется признаком или фактором выборки. Генеральная совокупность есть статистический аналог случайной величины, её объем N обычно велик, поэтому из неё выбирается часть данных, называемая выборочной совокупностью или просто выборкой

$$x_B = \{x_1, x_2, x_3, ..., x_n\} = \{x_i; i = 1, n\}, \quad x_B \subset x_{\Gamma}, n \le N.$$

Использование выборки для построения закономерностей, которым подчинена наблюдаемая случайная величина, позволяет избежать её сплошного (массового) наблюдения, что часто бывает ресурсоёмким процессом, а то и просто невозможным. Однако выборка должна удовлетворять следующим основным требованиям:

- выборка должна быть представительной, т.е. сохранять в себе пропорции генеральной совокупности,
- объём выборки должен быть небольшим, но достаточным для того, чтобы полученные результаты её анализа обладали необходимой степенью надёжности,
- данные в выборке не должны бать «засорены» грубыми измерениями, содержащими нетипично большие ошибки измерений.

Отметим, что в более строгом смысле выборку можно представить как случайную многомерную величину $\vec{X}_B = \{X_1, X_2, X_3, ..., X_n\} = \{X_i; i = 1, n\}$, у которой все компоненты X_i распределены одинаково и по закону распределения наблюдаемой случайной величины. В этом смысле выборочные значения x_B есть одна из реализаций величины $\vec{X}_B$.

Возможные значения элементов выборки $x_B = \{x_i; i = 1, n\}$, называются вариантами x_j выборки, причём число вариант m меньше, чем объём выбор-

ки n. Варианта может повторяться в выборке несколько раз, число повторения варианты x_j в выборке называется частотой варианты n_j. Причём, $n_1 + n_2 + ... + n_m = n$. Величина $w_j = n_j / n$ называется относительной частотой варианты x_j.

Упорядоченный по возрастанию значений набор вариант совместно с соответствующими им частотами называется вариационно-частотным рядом выборки:

$$V_{xn} = \{x_j, n_j; j = 1, m\}; \; V_{xw} = \{x_j, \nu_j; j = 1, m\}.$$

Ломаная линия, соединяющая точки вариационно-частотного ряда на плоскости (x, n) или (x, ν) называется полигоном частот.

Вариационно-частотный ряд имеет существенный недостаток, а именно, ненаглядность полигона в случае малой повторяемости вариант, например, при наблюдении непрерывного признака его повторяемость в выборке маловероятна. Более общей формой описания элементов выборки является гистограмма выборки.

Для построения гистограммы разобьём интервал значений выборки $R = x_{\max} - x_{\min}$ на m интервалов $h_j = (x_j, x_{j+1})$ длины $h = R / m$ с границами $x_j = x_{\min} + h \cdot (j - 1)$. Число элементов выборки x_B, попадающих в интервал, h_j называется частотой n_j интервала, кроме того вводятся следующие величины:

$\nu_j = n_j / n \;\sim$ относительная частота интервала,

$w_j = \nu_j / h_j \;\sim$ плотность относительной частоты интервала.

Совокупность интервалов, наблюдаемой в выборке случайной величины и соответствующих им частот, называется гистограммой выборки. Различаются гистограммы частот, относительных частот и плотности частоты и обозначаются соответственно:

$$H_{xn} = \{h_j, n_j; j = 1, m\}, \quad H_{x\nu} = \{h_j, \nu_j; j = 1, m\}, \quad H_{xw} = \{h_j, w_j; j = 1, m\}.$$

Для частот гистограммы выполнены следующие условия нормировки:

$$\sum_{j=1}^{m} n_j = n, \qquad \sum_{j=1}^{m} \nu_j = 1, \qquad \sum_{j=1}^{m} w_j h = 1$$

Число интервалов гистограммы m должно быть оптимальным, чтобы, с одной стороны, была достаточной повторяемость интервалов, а с другой стороны не должны сглаживаться особенности выборочной статистики. Рекомендуется значение $m \cong 1 + 3{,}2 \lg(n)$. На плоскости (x, n) гистограмма представляется ступенчатой фигурой.

Помимо полигона и гистограммы выборка характеризуется следующими *основными числовыми характеристиками*:

$\bar{x}_B = \frac{1}{n}\sum_{i=1}^{n} x_i$ ~ выборочное среднее;

$D_B = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_B)^2$ ~ выборочная дисперсия;

$\sigma_B = \sqrt{D_B}$ ~ выборочное среднеквадратическое отклонение;

$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x}_B)^2$ ~ исправленная выборочная дисперсия;

$S = \sqrt{S^2}$ ~ исправленное выборочное среднеквадратическое отклонение (выборочный стандарт).

Пусть, например, дана выборка полуденных температур месяца Май своим вариационно-частотным рядом с объёмом $n = 31$.

x_j	0	2	3	7	8	12	14	16	19	23	25	27	30
n_j	2	1	1	2	3	4	2	3	6	2	1	3	1

Полигон и гистограмма данной выборки приводятся ниже на рис.2.1.

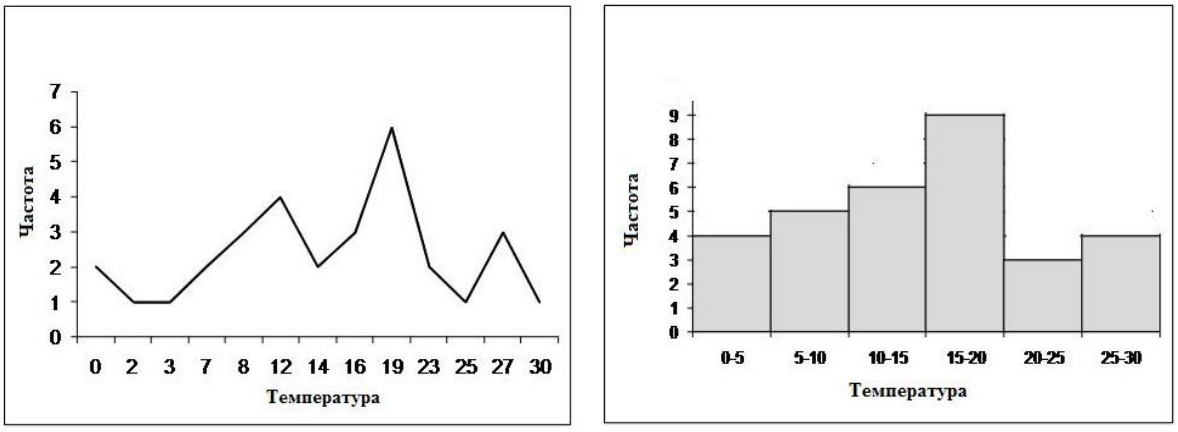


Рис. 2.1. Полигон и гистограмма частот выборки

Расчёт основных выборочных характеристик может быть легко проведен с помощью статистических функций приложения Excel-13 :

$\tilde{o}_{\hat{A}} = \frac{1}{n}\sum_{i=1}^{n} x_i = \tilde{N}\text{Ð}\text{Ç}\acute{I}\ \grave{A}\times(x_B) = 14{,}87$;

$D_{\hat{A}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_B)^2 = \ddot{A}\grave{E}\tilde{N}\ddot{I}\ .\tilde{A}(\tilde{o}_{\hat{A}}) = 60{,}31$;

$\sigma_{\hat{A}} = \sqrt{D_B} = \tilde{N}\grave{O}\grave{A}\acute{I}\ \ddot{A}\hat{I}\ \grave{O}\hat{E}\ddot{E}\hat{I}\ \acute{I}\ .\tilde{A}(\tilde{o}_{\hat{A}}) = 7.77$;

$S^2 = \frac{n}{n-1} D_B = \ddot{A}\grave{E}\tilde{N}\ddot{I}\ .\hat{A}(\tilde{o}_{\hat{A}}) = 62.32$;

$S = \sqrt{S^2} = \tilde{N}\grave{O}\grave{A}\acute{I}\ \ddot{A}\hat{I}\ \grave{O}\hat{E}\ddot{E}\hat{I}\ \acute{I}\ .\hat{A}(\tilde{o}_{\hat{A}}) = 7.89$.

Отметим, что все числовые характеристики выборки являются случайными величинами, поскольку получены по случайно взятой выборке. На элементах другой выборки наблюдений над той же случайной величиной X числовые характеристики в общем случае изменят свое значение.

Рассмотрим выборочные распределения нормальных выборок. Если наблюдаемая случайная величина X является нормальной, т.е $\tilde{O} = N(m,\sigma)$, где m - математическое ожидание, σ - среднеквадратическое отклонение, то случайная величина среднего выборочного $\overline{X}_B = \frac{1}{n}\sum_{i=1}^{n} X_i$ так же является нормальной $\tilde{O}_{\hat{A}} = N(m,\sigma/\sqrt{n})$. Здесь $\tilde{O}_i = N(m,\sigma)$ нормальные случайные величины, совпадающие с наблюдаемой величиной. Рассмотрим стандартные нормальные величины $\xi = N(0;1)$ в виде:

$$\xi_0 = \frac{X_B - a}{\sigma/\sqrt{n}}, \quad \xi_i = \frac{X_i - a}{\sigma}$$

и построим из них случайные величины Пирсона χ^2_n и Стьюдента t_n [4,8]:

$$\chi^2_{n-1} = \sum_{i=1}^{n}\xi_i^2 = \frac{1}{\sigma^2}\sum_{i}^{n}(X_i - a)^2 = \frac{nD_B}{\sigma^2} = \frac{n-1}{\sigma^2}S^2,$$

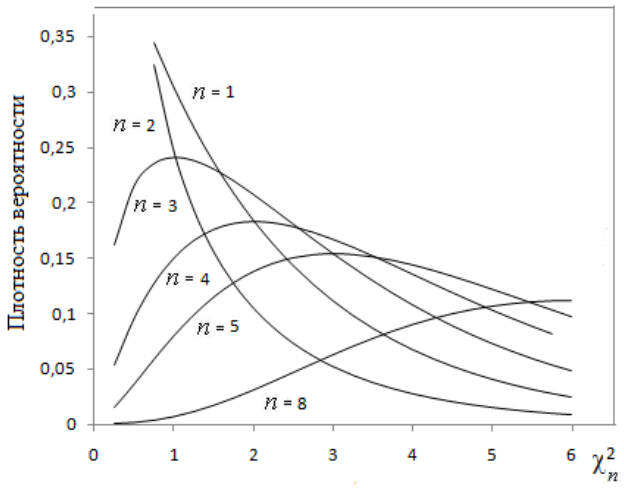
$$t_{n-1} = \frac{\xi_0}{\sqrt{\chi^2_n/(n-1)}} = \frac{X_B - a}{\sigma_B/\sqrt{n-1}} = \frac{X_B - a}{S/\sqrt{n}}.$$

Отсюда видно, что случайная величина выборочной дисперсии D_B распределена пропорционально «Хи-квадрат» случайной величине с *n-1* степенью свободы, а отклонение выборочного среднего от математического ожидания распределено пропорционально *t*-величине Стьюдента с *n-1* степенью свободы. При сравнении двух выборок объёмов n_1 и n_2 часто используется случайная величина Фишера [8] со степенями свободы n_1 и n_2:

$$F_{n_1,n_2} = \frac{\chi^2_{n_1}/n_1}{\chi^2_{n_2}/n_2}.$$

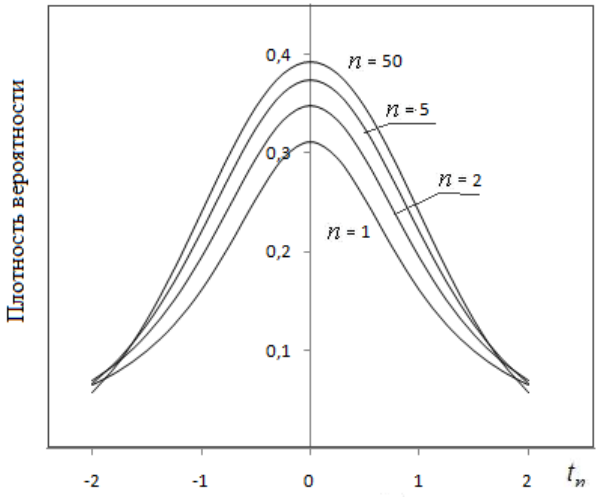
Распределения этих величин, как функций от стандартных нормальных величин, хорошо изучены и построены их функции распределения, обратного

распределения и плотности вероятности распределения. Ниже рис. 2.2.-2.4 представлены графики и функции Excel для их вычисления.



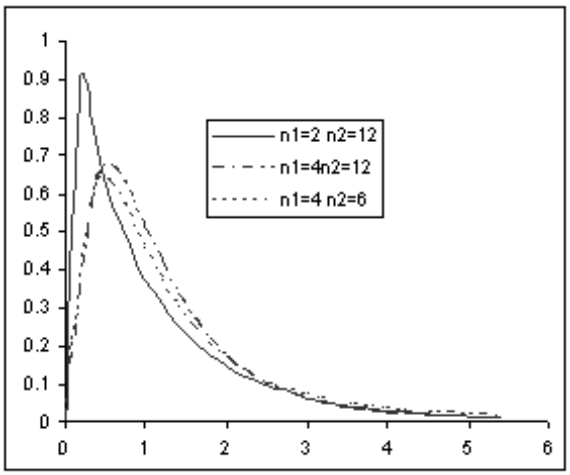
$$f(\chi_n^2) = ХИ2.РАСП(\chi^2, n, 0),$$
$$F(\chi_n^2) = ХИ2.РАСП(\chi^2, n, 1),$$
$$\chi_p^2 = ХИ2.ОБР(p, n),$$
$$\chi_\alpha^2 = ХИ2.ОБР.ПХ(\alpha, n)$$

Рис. 2.2. Функции распределения величины Пирсона



$$f(t_n) = СТЬЮДЕНТ.РАСП(t, n, 0),$$
$$F(t_n) = СТЬЮДЕНТ.РАСП(t, n, 1),$$
$$t_p = СТЬЮДЕНТ.ОБР(p, n),$$
$$t_\alpha = СТЬЮДЕНТ.ОБР.ПХ(\alpha, n)$$
$$t_{\alpha/2} = СТЬЮДЕНТ.ОБР.2Х(\alpha, n)$$

Рис. 2.3. Функции распределения величины Стьюдента



$$f(F) = F.РАСП(F, n1, n2, 0),$$
$$F_{\text{распр}}(F) = F.РАСП(F, n1, n2, 1),$$
$$F_p = F.ОБР(p, n1, n2),$$
$$F_\alpha = F.ОБР.ПХ(\alpha, n1, n2)$$

Рис. 2.4. Функции распределения величины Фишера

2.2 Статистические оценки

Пусть распределение наблюдаемой случайной непрерывной величины X (признак генеральной совокупности) задаётся функцией плотности вероятности $f_X(x,\theta)$, где θ параметр или параметры распределения. Допустим, что вид функции $f_X(x,\theta)$ известен или ограничен некоторым классом функций, а параметр θ неизвестен и должен быть оценён по выборке $x_B = \{x_i, n\} = \{x_1, x_2, ...x_n\}$, где n – объём выборки.

Точечной статистической оценкой параметров распределения или характеристик наблюдаемой случайной величины X называется построенная по данным выборки объема n величина:

$$\theta_n^* = \theta_n^*(x_1, x_2, ...x_n).$$

Например, статистическими оценками математического ожидания величины могут быть такие оценки: $m^* = \bar{x}_B$, $m^* = 0.5(x_{\min} + x_{\max})$ или $m^* = 0.75x_{\max}$.

Оценка θ_n^* является случайной величиной, т.к. зависит от случайной выборки. Для того, чтобы оценки, получаемые по данным различных выборок соответствовали истинному значению параметра θ, оценка должна удовлетворять следующим требованиям [8].

Оценка должна быть ***несмещенной***, т.е. её математическое ожидание должно совпадать с истинным значением параметра для любого объёма n

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

или хотя бы асимптотически несмещённой: $M(\theta_n^*) \xrightarrow[n\to\infty]{} \theta$.

Оценка должна быть ***состоятельной***, т.е. с ростом объёма выборки оценка должна сходиться по вероятности к истинному значению параметра:

$$P(|\theta_n^* - \theta| < \varepsilon) \xrightarrow[n\to\infty]{} 1 \quad \text{для любого } \varepsilon > 0.$$

Для состоятельности оценки достаточно выполнения следующего:

$$D(\theta_n^*) \xrightarrow[n\to\infty]{} 0.$$

Построенная оценка для использования на практике должна быть ***эффективной***, т.е. её дисперсия должна быть минимальной среди всех возможных оценок при фиксированном объёме выборки:

$$D(\theta_{nef}^*) = \min D(\theta_n^*).$$

Коэффициент эффективности оценки $k_{ef} = D(\theta_{nef}^*)/D(\theta_n^*)$ показывает степень эффективности оценки θ_n^*, если $k_{ef}(\theta_n^*) \xrightarrow[n\to\infty]{} 1$, то говорят об асимптотической эффективности оценки.

Отметим, что на практике не всегда удаётся удовлетворить всем перечисленным требованиям к оценке, но введённые свойства оценок позволяют проранжировать имеющиеся оценки по их качеству.

Как пример рассмотрим оценки математического ожидания $M(X)=m$ и дисперсии $D(X)=d$ наблюдаемой случайной величины X.
Построим точечные оценки:

$$m^* = \bar{x}_B, \qquad d^* = D_{\hat{A}}$$

и рассмотрим их свойства.

Поскольку можно вычислить, что для оценки m^* справедливо:

$$M(m^*)=m; \qquad D(m^*)=(d/n)\to 0 \quad \text{при } n\to\infty,$$

то из этого следует несмещённость и состоятельность оценки m^*.

Рассматривая же оценку d^*, можно получить что:

$$M(d^*)=\frac{n-1}{n}d \neq d \;;\; D(d^*)\approx \frac{1}{n}\xrightarrow[n\to\infty]{} 0.$$

Из чего следует состоятельность, но смещённость оценки d^*. Смещёность оценки здесь легко может быть исправлена, если рассмотрим оценку:

$$d^* = \frac{n}{n-1}D_{\hat{A}} = S^2.$$

Оценка $d^* = S^2$ является уже не только состоятельной, но и несмещённой, так как $M(d^*)=d$. Величина S^2 называется исправленной (несмещённой) выборочной дисперсией, а величина S - исправленным среднеквадратическим выборочным отклонением (выборочный стандарт).

В отличие от ***точечных*** оценок типа $\theta \approx \theta_n^*$ ***интервальные*** оценки задают интервал значений, где оцениваемый параметр находится с заданной вероятностью, т.е. это оценки типа $P(\left|\theta-\theta_n^*\right|\le\varepsilon)=\gamma$.

Надёжностью оценки (доверительной вероятностью) называется вероятность γ, с которой оцениваемый параметр находится в интервале:

$$\theta_n^* - \varepsilon_\gamma \le \theta \le \theta_n^* + \varepsilon_\gamma .$$

Полуширина доверительного интервала ε_γ называется точностью оценки, соответствующей надёжности γ. Для построения доверительного интервала (нахождения по γ величины ε_γ) необходимо знать закон распределения оценки случайной величины θ_n^*.

Пусть в выборке $x_B=\{x_i; i=1,n\}$ наблюдается нормальная случайная величина $X=N(m,\sigma)$ с неизвестными параметрами распределения m и σ.
Построим доверительный интервал для математического ожидания m:

$$\tilde{o}_{\hat{A}} - \varepsilon_\gamma \le m \le \tilde{o}_{\hat{A}} + \varepsilon_\gamma ,$$

принимая за точечную оценку m, величину $m^* = \tilde{o}_{\hat{A}}$ и учитывая, что величина $(\tilde{o}_{\hat{A}} - m)/(S/\sqrt{n}) = t_{n-1}$ имеет распределение Стьюдента с $n-1$ степенью свободы.

Решение уравнения $P(|\overline{x}_B - m| \le \varepsilon) = \gamma$ относительно ε при заданном значении γ эквивалентно решению уравнения:

$$Ð(\frac{|\overline{x}_Â - m|}{S/\sqrt{n}} < \frac{\varepsilon}{S/\sqrt{n}}) = \gamma, \text{ или } P(|t| < t_\gamma) = \gamma.$$

Его решение получим в виде $\varepsilon_\gamma = t_\gamma \cdot S/\sqrt{n}$, где $t_\gamma = t_{î\,áð}(1-\gamma, n-1) = ÑÒÜÐ\ ÄÅÍ\ Ò.ÐÀÑÏ\ .2Õ(1-\gamma, n-1)$ двухсторонняя квантиль Стьюдента (рис. 2.5).

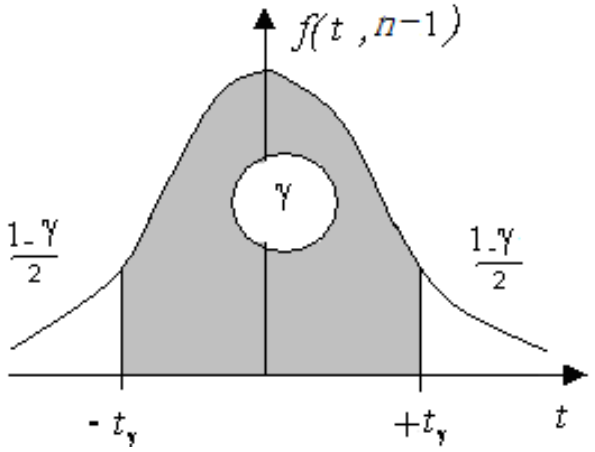


Рис. 2.5. Двухсторонняя квантиль Стьюдента

Построим теперь доверительный интервал для среднеквадратического отклонения σ:

$$S - \varepsilon_\gamma \le \sigma \le S + \varepsilon_\gamma.$$

Принимая за оценку σ величину $\sigma^* = S$ и учитывая, что величина $S^2(n-1)/\sigma^2 = \chi^2_{n-1}$, имеет χ^2-распределение с n - 1 степенью свободы. Решение уравнения $P(|S - \sigma| \le \varepsilon) = \gamma$ относительно ε при заданном параметре γ эквивалентно решению уравнения:

$$P(\chi^2_+ < \frac{S^2(n-1)}{\sigma^2} < \chi^2_-) = \gamma,$$

тогда получим его решение в виде $S\sqrt{\frac{n-1}{\chi^2_-}} < \sigma < S\sqrt{\frac{n-1}{\chi^2_+}}$, где величины $\chi^2_\pm = \chi^2_{î\,áð}(\frac{1\pm\gamma}{2}, n-1) = ÕÈ\ 2.Î\ ÁÐ.Ï\ Õ(\frac{1\pm\gamma}{2}, n-1)$ являются правосторонними «хи-квадрат» квантилями (рис. 2.6).

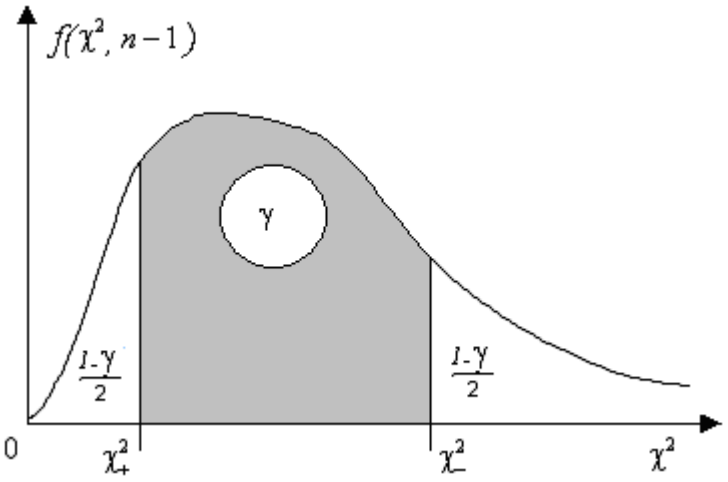


Рис. 2.6. Двухсторонняя «хи-квадрат» квантиль

Пример. Наблюдается выборка полуденных температур в Мае объёмом n =31 со средним выборочным значением $\tilde{o}_{\hat{A}} = 14{,}87$ и несмещённой дисперсией $S^2 = 62.32$. Построить доверительные интервалы для неизвестного математического ожидания m и среднеквадратического отклонения σ при надёжности $\gamma = 0{,}95$.

Исправленное выборочное среднеквадратическое отклонение $S = 7.89$.

Через обратное распределение Стьюдента находим $t_\gamma = \tilde{N}\grave{O}\ddot{U}\acute{I} \ddot{A}\mathring{A}\acute{I} \grave{O}.\hat{I} \acute{A}\eth.2\tilde{O}(1-\gamma, n-1) = 2{,}042$, тогда $\varepsilon_\gamma = 2.042 \cdot 7{,}89 / \sqrt{31} = 2.894$ и тогда доверительный интервал для математического ожидания m будет:

$$14.87+2.894 < m < 14.87+2.894 \text{ или } 11.976 < m < 17.674.$$

Для построения доверительного интервала среднеквадратического отклонения через обратное распределение «Хи-квадрат» находим

$\chi^2_- = \tilde{O}\grave{E}\ 2.\hat{I}\ \acute{A}\eth.\ddot{I}\ \tilde{O}(\frac{1-\gamma}{2}, n-1) = 49.98$ $\chi^2_+ = \tilde{O}\grave{E}\ 2.\hat{I}\ \acute{A}\eth.\ddot{I}\ \tilde{O}(\frac{1+\gamma}{2}, n-1) = 16.79$, тогда:

$$6.305 = 7.89\sqrt{\frac{31-1}{49.98}} < \sigma < 7.89\sqrt{\frac{31-1}{16.74}} = 10.546.$$

2.3. Проверка статистических гипотез

Имея дело со случайными величинами, в различных областях человеческой деятельности часто приходится высказывать предположения о виде распределения случайной величины или о значениях её параметров. Эти предположения строятся с целью прогнозирования поведения случайной величины и принятия решений в условиях неопределённости.

Статистической гипотезой называется любое предположение о виде распределения случайной величины $f_X(x, \theta)$ или/и о значении неизвестных параметров распределения θ:

$H = \{X \sim f_X(x, \theta); \theta = \theta_0\}$ – статистическая гипотеза.

Высказанная статистическая гипотеза должна быть проверена по результатам наблюдений (измерений) случайной величины [11], в результате чего, гипотеза принимается или отвергается с определённой степенью риска совершить ошибку. Примером статистической гипотезы может быть предположение о том, что наблюдаемая в выборке случайная величина является нормальной с определёнными значениями параметров:

$$H = \{X = N(m, \sigma); m = \bar{x}_B; \sigma = S\}.$$

Выдвинутая статистическая гипотеза *H* должна быть проверена. Как и в любой другой науке, критерием её проверки является опыт, т.е. наблюдение (измерение) случайной величины. Критерий проверки должен отвергать или принимать гипотезу по результатам наблюдения. При этом могут быть совершены ошибки двух родов [6]:

1. Отвергнута верная гипотеза с вероятностью α,

2. Принята не верная гипотеза с вероятностью β.

Исключить эти ошибки полностью невозможно («не ошибается тот, кто ничего не делает»), но их можно постараться минимизировать. Учитывая сказанное, при построении критерия проверки статистической гипотезы необходимо сначала задаться допустимым уровнем риска на совершение ошибки 1 рода, как наиболее значимой, а затем минимизировать ошибки 2 рода.

Пусть необходимо проверить гипотезу $H_0 = \{X \leftrightarrow f_X(x, \theta)\}$, помимо основной гипотезы H_0 («нулевой») рассмотрим ещё одну или несколько альтернативных гипотез $Í_1, H_2, H_3, ..$ каждая из которых противоречит основной. Построим критерий, однозначно принимающий или отвергающий проверяемую гипотезу по полученной в наблюдении за случайной величиной X выборке $x_B = \{x_1, x_2, ...x_n\}$ объёма *n*. Критерий проверки гипотезы состоит из двух составляющих:

Во-первых, в качестве критерия принимается некоторая случайная величина $\hat{E}$, с известным распределением при условии справедливости основной $f_K(k / H_0)$ и хотя бы частично известным для альтернативных гипотез $f_K(k / H_j)$ j=1, ..*m*. Кроме того, значения критерия должны быть вычисляемы по наблюдаемой выборке x_B, т.е. $k_{íàá} = k(x_i)$.

Во-вторых, строится решающее правило для критерия проверки, согласно которому гипотеза будет приниматься или отвергаться. Для этого, назовем критической областью критерия те значения величины $\hat{E}$, при которых гипотеза отвергается. Критическую область будем обозначать K_{kr}. Тогда решающее правило критерия проверки будет следующим:

$k_{íàá} \in \hat{E}_{kr} \Rightarrow Í_0$ отвергается (по наблюдаемой выборке),

$k_{íàá} \notin \hat{E}_{kr} \Rightarrow Í_0$ принимается (нет оснований отвергать гипотезу).

Точки значения критерия $\hat{E}$, где критическая область критерия проверки K_{kr} отделяется от области принятия гипотезы, называются критическими точками критерия k_{kr}. Как построить критическую область критерия?

Принцип максимального правдоподобия утверждает, что наблюдаемые события имеют большую вероятность и наоборот, маловероятные события ненаблюдаемые. Согласно этому принципу наблюдаемое значение критерия $k_{íàá}$ должно иметь в рамках проверяемой гипотезы большую вероятность. В противном случае, если вероятность наблюдаемой величины мала, проверяемую гипотезу нужно отвергать в пользу иных альтернативных гипотез.

Зададимся вероятностью α ошибки 1-го рода, как наиболее значимой. Исключить такую ошибку при проверке гипотезы невозможно ($\alpha \neq 0$), на практике обычно эту вероятность задают достаточно малой величиной $\alpha = 0{,}05$; $\alpha = 0{,}025$; $\alpha = 0{,}005$ и называют уровнем значимости гипотезы.
Если из условия

$$P(k \in K_{kr}) = \int_{K_{kr}} f_K(k / H_0)dx = \alpha,$$

можно определить критические точки k_{kr} однозначно, то задача построения критической области критерия решена. В противном случае, когда ещё остаётся свобода выбора критических точек, рассмотрим влияние альтернативных гипотез. Поскольку величина β_j - есть вероятность принять неверную гипотезу H_0 при условии справедливости альтернативной гипотезы H_j, то

$$\int_{K_{kr}} f_K(k / H_j)dx = 1 - \beta_j$$

есть вероятность правильного отбрасывания H_0 при условии справедливости H_j и её называют мощностью критерия по отношению к альтернативной гипотезе H_j. Поэтому при заданном уровне значимости α, критическую область критерия нужно строить так, чтобы мощность критерия была максимальной $(1-\beta_j) \Rightarrow \max$ по отношению ко всем альтернативным гипотезам.

Таким образом, критическими точками критерия являются квантили его распределения, определенные согласно уровню значимости проверяемой гипотезы.

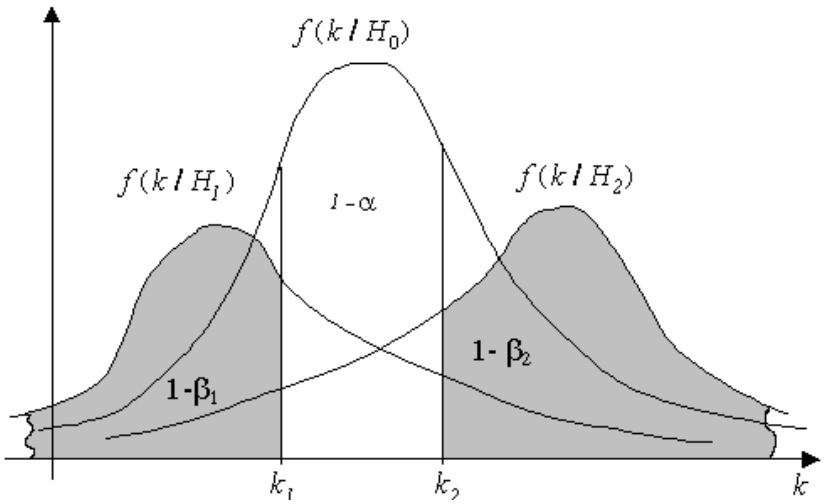


Рис. 2.7. Двухсторонняя критическая область критерия $K_{kr} = \{k > k_2, k < k_1\}$ при наличии двух альтернативных гипотез H_1, H_2

На рис. 2.7 приведена графическая интерпретация алгоритма построения критической области одномерного критерия. Видим, что структура критической области зависит от наличия альтернативных гипотез и их «расположения» относительно основной.

Рассмотрим примеры.

Критерий Смирнова-Граббса. Рассмотрим проблему отсева грубых ошибок при измерении нормальной случайной величины. Пусть мы имеем нормальную выборку наблюдений $x_B = \{\tilde{o}_i; n\}$ объёмом n, а проверяемой гипотезой является гипотеза о не грубой ошибке при измерении элемента $\tilde{o}_j$ этой выборки. Тогда $H_0 = \{X = N(m,\sigma),\ \tilde{o}_j \in \tilde{o}_{\hat{A}}\}$, $H_1 = \bar{H}_0$. Критерием для проверки гипотезы является величина Стьюдента

$$K = \frac{\left|x_j - \overline{x}_B\right|}{S} = t_n .$$

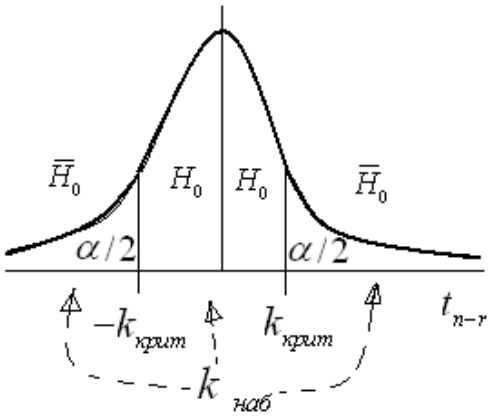
Вычисляя значение $k_{í àá}$ и критическую точку при заданном уровне значимости α проверяемой гипотезы $k_{êðèò} = ÑÒÜÞ ÄÅÍ ÒÎ ÁÐ.2Õ(\alpha, n-1)$ можно судить о грубости данного измерения. Обычно на грубость измерения проверяются крайние точки наблюдений (максимальная и минимальная). Проверим на грубость измеренную максимальную температуру в рассмотренной выше выборке майских температурных измерений.

$x_j = 30, \overline{x}_B = 14.87, S = 7.89, \alpha = 0.1$, $k_{í àá} = \dfrac{\left|30 - 14.87\right|}{7.89} = 1.918$, $k_{êðèò} = 1.697$

Видим, что при значимости проверяемой гипотезы в 10% критерий отклоняет её в пользу гипотезы H_1 о грубости этого измерения. Таким образом, это измерение грубое и его лучше убрать из выборки. Вывод критерия зависит от точности измерения (её объёма n) и значимости гипотезы, то есть риска ошибиться при отклонении верной гипотезы. Так, если уровень значимости ги-

потезы повысить до 5%, то $k_{êðèò} = 2.042$, то измерение уже не является грубым.

Критерий Стьюдента о значимости измеренной величины. В статистическом анализе очень часто используются критерии о значимости оценок различных величин, построенных по выборке. Проверяемой гипотезой является гипотеза о том, что истинная теоретическая величина u равна нулю $H_0 = \{u = 0\}$, а в наблюдениях ее выборочный аналог u_B отличен от нуля. Действительно ли наблюдаемое значение не нулевое (значимое), или это произошло случайно на рассматриваемой выборке? Для ответа на этот вопрос очень часто в дальнейшем мы будем использовать критерий Стьюдента рис. 2.8 в виде:



$$K = \frac{|u_B|}{S_u} = t_{n-r},$$

$$k_{êðèò} = ÑÒÜÞÄÅÍ ÒÎ ÁÐ2Õ(\alpha, n-r)$$

Рис. 2.8. Критерий Стьюдента проверки значимости величины

Здесь u_B, S_u статистическая оценка и её несмещённая ошибка, r количество степеней свободы выборки, потерянных при построении оценки. Для удобства часто вводится понятие жёсткости критерия

$$G(\alpha) = \frac{k_{íàáë}}{k_{êðèò}}.$$

Значимость проверяемой оценки имеет место быть при жёсткости $G > 1$, когда проверяемая гипотеза о нулевом значении теоретической величины отвергается.

Например, в качестве проверяемой величины часто используется выборочный коэффициент корреляции между двумя выборками x_B, y_B одинакового объёма.

$$r_{xy} = \frac{\overline{x \cdot y} - \bar{x} \cdot \bar{y}}{\sigma_x \cdot \sigma_y} = ÊÎ ÐÐÅË\left(x_B; y_B\right), \quad \overline{x \cdot y} = \frac{1}{n}\sum_{i=1}^{n} x_i \cdot y_i = ÑÓÌ\ Ì\ Ï ÐÎ ÈÇÂ\left(x_B; y_B\right)/n$$

Критерием является следующая величина Стьюдента:

$$\hat{E} = \frac{r_{xy}\sqrt{n-2}}{\sqrt{1-r_{xy}^2}} = t_{n-2}.$$

3. Многомерные статистические данные

Измерительные данные, с которыми работает инженер-исследователь или аналитик в процессах проектирования, производства, эксплуатации и мониторинга различных технических, экологических, социально-экономических систем редко бывают одномерными. Обычно при исследовании объекта или множества объектов измеряется несколько параметров объекта. Таким образом формируется многомерный статистический набор данных. При строительстве и эксплуатации зданий и сооружений могут быть измерены и запротоколированы множество различных параметров (рис. 3.1).

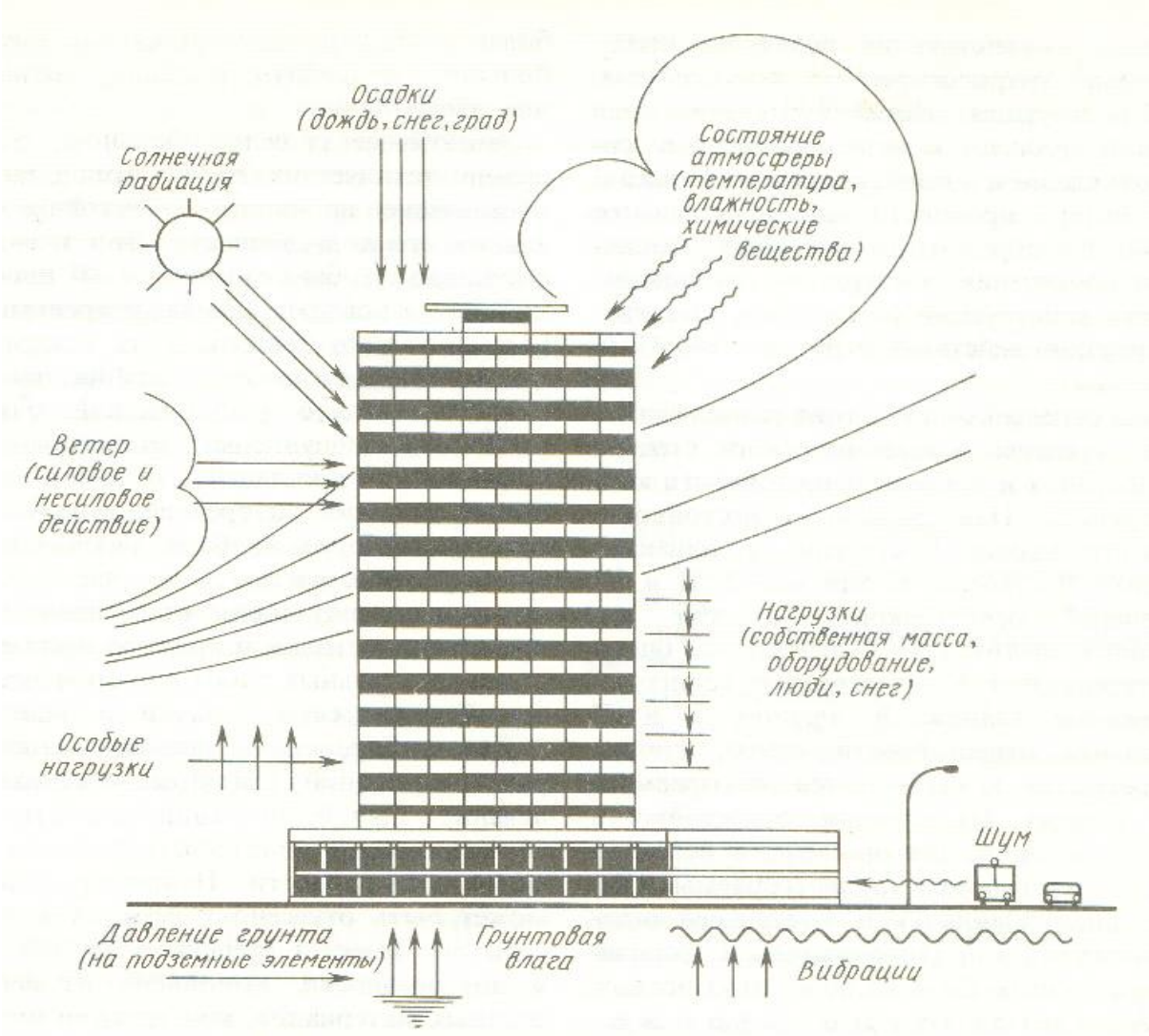


Рис. 3.1. Факторы влияния на здание

Измеряемые величины в большинстве случаев являются случайными как по своей природе, так и за счёт ошибок измерения

$$x = x_0 + \delta_x + \Delta_x,$$

где x_0 - истинное или среднее значение величины, δ_x - флуктуация измеряемой величины, Δ_x - ошибка измерительного прибора и измеряющего субъекта. Виды измерений разнообразны и классифицируются по множеству признаков (рис. 3.2).

Рис. 3.2.Виды измерений

Приведём несколько примеров наборов статистических данных, как документально оформленных измерений.

ООО «Базис»
Испытательная строительная лаборатория
Аттестационное свидетельство №02-1218 до 30.07.2011
Адрес: 140180, Московская область Раменский район, п. Быково, ул. Театральная, стр. 8
Тел./факс: 8(495) 221-68-02, 8 (495) 556-62-46, 556-10-62

Пртокол испытаний №06/11-09

Заказчик ООО «Еврокомплектстрой» Объект, адрес _____
Испытуемый образец кубов фибробетона ГОСТ 10180-90 дата поступления пробы 26.10.09
Испыт. оборудование ИП-1А-1000 Дата поверки VI-2009 Измерение, наблюдения получ. рез-тов -
_____ Погрешность измерений 01,мм

Дата изготовления образца	Состав пескобетона, кг						Дата испытания		Осадка конуса, см	Плотность, г/дм3	Прочность на сжатие, МПа
	Цемент М400Д5	Песок II Кл, МКр3,0	Вода	Эластобетон-А	Фибра	Пено-гаситель	Возраст образца				
26.10.09	1.0	2.0	0.37	0.038	нет	нет	02.11.09	7сут	8.5	2143	40.0
27.10.09	1.0	2.0	0.40	0.038	0.005	нет	03.11.09	7сут	8.7	2184	41.6
28.10.09	1.0	2.0	0.5	нет	нет	нет	05.11.09	8сут	8.5	2148	21.4
29.10.09	1.0	2.0	0.37	0.038	нет	0.001	05.11.09	7сут	9.0	2260	41.8
03.11.09	1.0	2.0	0.4	0.038	0.005	0.001	10.11.09	7сут	8.2	2242	49.3

Начальник ИСЛ ООО «Базис» ИСПЫТАТЕЛЬНАЯ СТРОИТЕЛЬНАЯ ЛАБОРАТОРИЯ ООО «БАЗИС» Лебедева Б.А.

ООО "Оргтехстрой"
603074, г. Н.Новгород,
ул. Народная, 50
тел. 241-50-73
Отдел лабораторных испытаний
строительных материалов и
конструкций

РЕЗУЛЬТАТЫ

Испытаний образцов бетона **плит бетонных фасадных** на морозостойкость **F100** методом многократного замораживания и оттаивания
«Заказчик» ООО «Новабрик - Восток»
Испытание на морозостойкость проводилось в соответствии с требованиями ГОСТ 10060.0-95, 10060.2-95 (III метод) и ГОСТ 6927-74.
Образцы изготовлены «Заказчиком».

Исходные данные контрольных и основных образцов			Результаты испытаний образцов													
			Контрольных			Основных, после итоговых испытаний										
Маркировка образца		Размер образца, см	Дата испытания	Прочность на сжатие в насыщенном состоянии кгс/см^2	Средняя прочность кгс/см^2	Масса образца г до начала испытания	Средняя масса г	Дата итогового испытания	Число циклов с начала испытания	Прочность на сжатие кгс/см^2	Средняя прочность кгс/см^2	Масса образца г после испытания	Средняя масса г	Уменьшение средней массы,%	Изменение средней прочности, %
Контрольные	1	7,0x7,0x7,0	14.02. 2014 г.	302	318									Снижения массы нет. Δm=0 %	Снижения прочности нет. ΔR= +19 %
	2	7,0x7,0x7,0		~~245~~											
	3	7,0x7,0x7,1		272											
	4	7,1x6,9x7,0		331											
	5	6,9x7,0x7,0		318											
	6	6,9x7,0x7,0		320											
Основные	7	7,0x7,0x7,0				855	862	19.02. 2014 г.	3 цикла	345	380	855	862		
	8	7,0x7,0x7,1				870				415		870			
	9	7,0x7,1x7,0				880				325		880			
	10	7,0x7,1x7,0				875				~~304~~		875			
	11	7,0x7,0x7,0				840				412		840			
	12	6,9x7,0x7,0				850				348		850			

Заключение: Данная партия бетона плит бетонных фасадных соответствует требуемой марке по морозостойкости F100.

Начальник лаборатории И.П. Миронова

ИСПЫТАТЕЛЬНАЯ ЛАБОРАТОРИЯ ООО "Оргтехстрой"

Состав	Расход цемента, кг/м³ бетона	Колич. добавки, мас.% от расхода цемента	Соотношение хлорида кальция и хлорида натрия	Содержание хлорида кальция и хлорида натрия кг на 1м³ бет. смеси	Прочность бетона на сжатие, МПа		
					после ТВО	через 28 сут	через 28 сут Нормального хранения
1 контр.	400				28,3	42,7	27,5
2прототот	400	1,8	1,01:1,0	7,2	25,0	37,0	-
3	400	2,8	1,01:1,0	11,2	35,0	53,3	38,3
4	400	3,0	1,01:1,0	12,0	36,0	45,5	42,0
5	350	-			27,5	41,1	27,5
6	350	1,0	1,25:1	3,5	28,7	43,2	30,2
7	350	2,0	1,25:1	7,0	33,0	46,5	33,2
8	350	3,0	1,25:1	10,5	32,2	46,2	32,2
9	350	4,0	1,25:1	14,0	28,4	44,9	35,1
10	350	5,0	1,25:1	17,5	27,8	44,0	34,5
11	450	2,0	1,4:1	9,0	43,0	59,8	43,6
12	450	3,0	1,4:1	13,5	35,2	-	-

Многомерность статистических данных состоит в том, что у каждого наблюдаемого объекта $A_i(X_1, X_2, X_3, ..., X_m)$, измеряется (фиксируется) несколько величин-факторов $X_j \quad j=\overline{1,m}$. Измерения могут проводиться как одновременно по *n* однотипным объектам (пространственные ряды данных), так и *n* измерений одного объекта в разные моменты времени (временные ряды данных) рис.3.3.

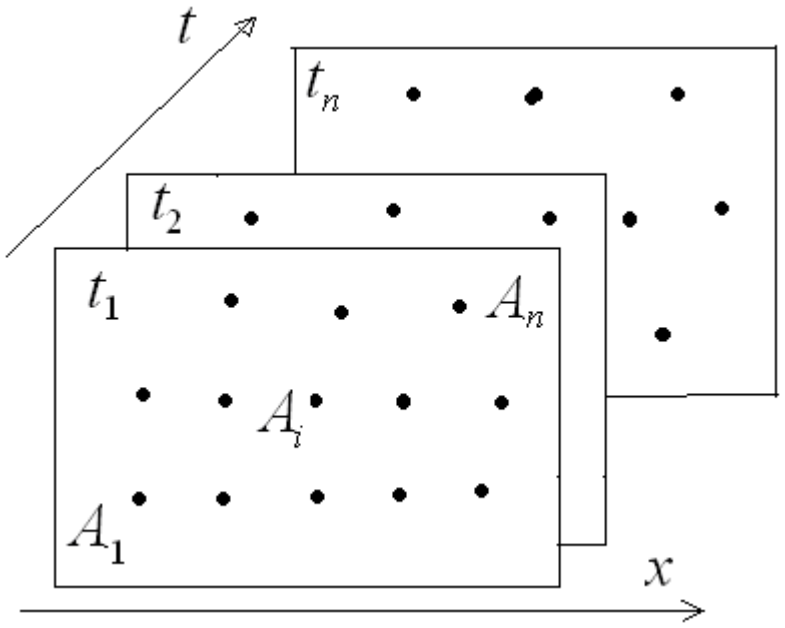


Рис. 3.3. Пространственные и временные ряды данных образуют куб данных

Каждый объект, в своём ряду данных, представляется вектором измерений $x=(x_1, x_2, ..., x_j, ..., x_m) \quad j=\overline{1,m}$. Объединим все измерения ряда в матрицу измерений.

$$X=\begin{pmatrix} x_{11} & x_{12} & x_{13} & ... & x_{1m} \\ x_{21} & x_{22} & x_{23} & ... & x_{2m} \\ x_{31} & x_{32} & x_{33} & ... & x_{3m} \\ ... & ... & ... & ... & ... \\ x_{n1} & x_{n2} & x_{33} & ... & x_{nm} \end{pmatrix}; \quad \begin{matrix} i=\overline{1,n} \\ j=\overline{1,m} \end{matrix}; \quad X=\begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ ... \\ x_n \end{pmatrix}; \quad x_i=(x_{i1}, x_{i2}, ..., x_{ij}, ..., x_{im})$$

Используя все измерения по *n* объектам, можем вычислить числовые характеристики по каждому измеримому фактору.

$$\begin{pmatrix} \overline{x}_1 & \overline{x}_2 & \overline{x}_3 & ... & \overline{x}_m \\ D_{x1} & D_{x2} & D_{x3} & ... & D_{xm} \\ \sigma_{x1} & \sigma_{x2} & \sigma_{x3} & ... & \sigma_{xm} \\ s_{x1} & s_{x2} & s_{x3} & ... & s_{xm} \end{pmatrix} \begin{matrix} =\tilde{N}\text{Ð}\text{Ç}\grave{A}\times \\ =\ddot{A}\grave{E}\tilde{N}\ddot{I}\ .\tilde{A} \\ =\tilde{N}\grave{O}\grave{A}\acute{I}\ \ddot{A}\hat{I}\ \grave{O}\hat{E}\ddot{E}\ .\tilde{A} \\ =\tilde{N}\grave{O}\grave{A}\acute{I}\ \ddot{A}\hat{I}\ \grave{O}\hat{E}\ddot{E}\ .\hat{A} \end{matrix}$$

Зная средние значения $\overline{x}_j$ и среднеквадратические отклонения σ_{x_j} по каждому измеримому фактору, проведём центрирование и нормирование переменных $x'_{ij}=\dfrac{x_{ij}-\overline{x}_j}{\sigma_{x_j}}$ и тем самым приведём матрицу измерений к стандартному

виду, в котором $\bar{x}'_j = 0, D_{x_j} = \sigma'_{x_j} = 1, s'_{x_j} = \sqrt{n/(n-1)}$. Помимо единого масштаба для всех измеряемых факторов, такой вид матрицы измерения, как увидим далее, позволяет упростить ряд статистических формул. Поэтому в дальнейшем будем пользоваться именно стандартной формой матрицы измерений, а штрихи будем отпускать. При необходимости всегда можно пересчитать все получаемые величины в реальный масштаб по формуле $x_{ij} = \bar{x}_j + \sigma_{x_j} x'_{ij}$.

Помимо преобразования в стандартную форму, рекомендуется проверить измерения на грубые ошибки согласно критерию Смирнова-Греббса [9-10].

Рассмотрим пример многомерных статистических данных, которые будем анализировать во всех последующих главах. Пример состоит в анализе данных об *n=11* земельных участках, проданных на рынке в течение года. Известны данные о следующих *m=4* факторах участка:

x_1 - урожайность участка (кг/сотка)

x_2 - экспертная оценка уровня инфраструктуры участка,

x_3 - экспертная оценка уровня экологии участка,

x_4 - признак принадлежности участка к землям населённых пунктов,

y - цена проданного на рынке участка (руб/сотка).

Реальньный масштаб

	x_1	x_2	x_3	x_1	y
X=	100	2	5	0	200
	90	2	4	1	1000
	50	1	7	0	100
	70	5	1	0	1500
	120	4	2	1	2500
	160	1	5	0	50
	70	2	3	1	900
	30	3	4	0	170
	150	3	7	0	80
	90	1	3	0	110
	30	6	1	1	3000
срзнач	87.273	2.727	3.818	0.364	873.636
дисп	1728.926	2.562	3.967	0.231	1001295.9
ско	41.580	1.601	1.992	0.481	1000.648
стандоткл	43.610	1.679	2.089	0.505	1049.488

Пересчитаем данные по 11 участкам в стандартную форму путём центрирования и нормирования факторов, а также проверим засорённость данных грубыми ошибками измерений [11].

Стандартный масштаб

	x_1	x_2	x_3	x_1	y
$X =$	0.306	-0.454	0.593	-0.756	-0.673
	0.066	-0.454	0.091	1.323	0.126
	-0.896	-1.079	1.598	-0.756	-0.773
	-0.415	1.420	-1.415	-0.756	0.626
	0.787	0.795	-0.913	1.323	1.625
	1.749	-1.079	0.593	-0.756	-0.823
	-0.415	-0.454	-0.411	1.323	0.026
	-1.377	0.170	0.091	-0.756	-0.703
	1.509	0.170	1.598	-0.756	-0.793
	0.066	-1.079	-0.411	-0.756	-0.763
	-1.377	2.045	-1.415	1.323	2.125
срзнач	0.000	0.000	0.000	0.000	0.000
дисп	1.000	1.000	1.000	1.000	1.000
ско	1.000	1.000	1.000	1.000	1.000
стандоткл	1.049	1.049	1.049	1.049	1.049
Gmax =	0.748	0.875	0.684	0.566	0.909
Gmin =	0.589	0.462	0.605	0.323	0.352

Пересчитанные данные имеют стандартные параметры и не имеют грубых ошибок измерения по уровню в 5%. Последнее видно из того, что жёсткость критерия Смирнова-Граббса нигде не превышает единицы, как по максимальным отклонениям, так и по минимальным отклонениям всех факторов.

4. Задачи корреляционного анализа

Задачей корреляционного анализа является определение статистической зависимости между наблюдаемыми величинами. Зависимость между величинами X_j и X_k определяется империческим коэффициентом парной корреляции

$$r_{jk} = \frac{\overline{x_j \cdot x_k} - \overline{x}_j \cdot \overline{x}_k}{\sqrt{D(x_k) \cdot D(x_k)}},$$

поскольку матрица измерений нормирована, то $r_{jk} = \overline{x_j \cdot x_k} = \frac{1}{n}\sum_{i=1}^{n} x_{ji} x_{ik}$. Все парные коэффициенты корреляции образуют симметричную корреляционную матрицу R:

$$R = (r_{jk}) = \begin{pmatrix} r_{11} & r_{12} & \dots & r_{1m} \\ r_{21} & r_{22} & \dots & r_{2m} \\ \dots & \dots & \dots & \dots \\ r_{m1} & r_{m2} & \dots & r_{mm} \end{pmatrix} = \frac{1}{n} \begin{pmatrix} \overline{x_1 \cdot x_1} & \overline{x_1 \cdot x_2} & \dots & \overline{x_1 \cdot x_m} \\ \overline{x_2 \cdot x_1} & \overline{x_2 \cdot x_2} & \dots & \overline{x_2 \cdot x_m} \\ \dots & \dots & \dots & \dots \\ \overline{x_m \cdot x_1} & \overline{x_m \cdot x_2} & \dots & \overline{x_m \cdot x_m} \end{pmatrix} = \overline{X \cdot X^T}.$$

Значения $|r_{jk}| \approx 0$ говорят о малой зависимости X_j и X_k наблюдаемых величин и напротив значения $|r_{jk}| \approx 1$ говорят о сильной (почти линейной) зависимости этих величин. Для более строгого определения зависимостей величин воспользуемся критерием Стьюдента с уровнем значимости α:

$$\frac{r_{jk}\sqrt{n-2}}{\sqrt{1-r_{jk}^2}} = t_{n-2}, \quad G_{jk} = \frac{t_{íàáë,jk}}{t_{êðèò}}, \quad t_{êðèò} = ÑÒÜÞ\ Ä.Î\ ÁÐ.2Õ(\alpha, n-2).$$

Величина G_{jk} показывает жёсткость корреляционной зависимости, а при $G_{jk} > 1$ эта зависимость является значимой по уровню α.

Более строгий корреляционный анализ многомерных данных проводится при помощи частных (очищенных) коэффициентов корреляции. Дело в том, что в многомерных данных парная корреляция двух переменных может быть установлена не по причине их зависимости между собой, а из-за их зависимости от третьей переменной. Частный коэффициент корреляции для переменных X_j и X_k по отношению к переменной X_l вычисляется так:

$$r'_{jk,l} = \frac{r_{jk} - r_{jl} \cdot r_{kl}}{\sqrt{(1-r_{jl}^2)(1-r_{kl}^2)}}.$$

В общем случае частный коэффициент корреляции, очищенный от влияния всех остальных переменных, вычисляется по формулам:

$$r'_{jk,l_1 l_2 \dots l_{m-2}} = \frac{-c_{jk}}{\sqrt{c_{jj} c_{kk}}}, \quad l \neq j, l \neq k, \quad C = (c_{jk}) = R^{-1}.$$

Числовой пример (часть 2)

Рассмотрим числовой пример для рассмотренной в части 1 матрицы измерений X, приведённой на странице 35 к стандартному виду и расширенной вектором измерений y. Матрица парных корреляций и их Стьюдентовской жёсткости будут такими:

$R =$

1.000	-0.339	0.367	-0.178	-0.286
-0.339	1.000	-0.671	0.365	0.823
0.367	-0.671	1.000	-0.500	-0.774
-0.178	0.365	-0.500	1.000	0.738
-0.286	0.823	-0.774	0.738	1.000

$G =$

9.999	-0.442	0.484	-0.221	-0.365
-0.442	9.999	-1.111	0.481	**1.777**
0.484	-1.111	9.999	-0.708	**-1.500**
-0.221	0.481	-0.708	9.999	**1.339**
-0.365	**1.777**	**-1.500**	**1.339**	9.999

Видим, что связь меду переменными X_2 и X_3 значима по уровню $\alpha = 0.05$. Переменная Y жёстко коррелирует с переменными X_2, X_3, X_4, что говорит о её зависимости от этих переменных.

Частные парные коэффициенты корреляции переменных X_j, очищенные от остальных переменных таковы:

$R' =$

1.000	0.214	-0.244	0.127	-0.168
0.214	1.000	-0.064	0.635	-0.810
-0.244	-0.064	1.000	-0.162	0.444
0.127	0.635	-0.162	1.000	-0.804
-0.168	-0.810	0.444	-0.804	1.000

$G' =$

9.999	0.219	-0.252	0.128	-0.171
0.219	9.999	-0.064	0.822	-1.383
-0.252	-0.064	9.999	-0.165	0.497
0.128	0.822	-0.165	9.999	-1.353
-0.171	-1.383	0.497	-1.353	9.999

Как видно, корреляционная связь меду переменными X_2 и X_3 уменьшилась, но возросла между X_2 и X_4, оставаясь не значимой. Связи переменной Y с другими переменными изменились, но значимость сохранилась. В расчётах учтено, что $t_{êðèò} = \tilde{N}\grave{O}\ddot{U}\acute{E}\ \ddot{A}.\hat{I}\ \acute{A}\text{Ð}.2\tilde{O}(\alpha, n-2) = 2.262$, а

$R^{-1} =$

1.212	0.548	-0.443	0.272	-0.649
0.548	5.404	-0.245	2.868	-6.596
-0.443	-0.245	2.729	-0.521	2.572
0.272	2.868	-0.521	3.777	-5.472
-0.649	-6.596	2.572	-5.472	12.271

.

5. Задачи регрессионного анализа

Если задачей корреляционного анализа является установление факта зависимости наблюдаемых случайных величин, то задачей регрессионного анализа является установление вида этой зависимости. Выделим из наблюдаемых величин случайную величину Y и постараемся объяснить её значения и свойства через значения других величин X_j. Функционально такую зависимость будем описывать регрессионной моделью:

$$Y = g(X_1, X_2, ..., X_m) + \varepsilon$$

Будем величину Y будем называть объясняемой, а величины X_j, ãäå $j = \overline{1,m}$ объясняющими. Регрессионную часть $\hat{Y} = g(X_1, X_2, ..., X_m)$ назовём объяснённой частью, а величину ε - необъяснённой (специфической, остаточной) частью объясняемой величины. Потребуем в модели для остаточного члена выполнения следующих условий:

А) $M(\varepsilon) = 0$. Это обеспечивает выполнение условия $M(Y) = M(\hat{Y})$.

Б) $D(\varepsilon) \Rightarrow \min\limits_{g(...)}$. Это условие качественной модели регрессии.

Важной проблемой регрессионного анализа является проблема спецификации модели, состоящая в определении состава объясняющих переменных для выбранной объясняемой переменной.

5.1 Линейная среднеквадратическая регрессия

Построим линейную регрессионную модель в виде:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m + \varepsilon .$$

Коэффициенты теоретической регрессии $\beta_j, j = \overline{1,m}$ необходимо определить из выше приведённых условий А и Б, на основе наблюдательных статистических данных, собранных в матрицу измерений X и измеренный вектор Y объёма n.

$$X = \begin{pmatrix} x_{11} & x_{21} & ... & x_{m1} \\ x_{12} & x_{22} & ... & x_{m2} \\ ... & ... & ... & ... \\ x_{1ò} & x_{2n} & ... & x_{mn} \end{pmatrix} = (x_1, x_2, ...x_m); \quad y = \begin{pmatrix} y_1 \\ y_2 \\ ... \\ y_n \end{pmatrix}, \quad x_j = \begin{pmatrix} x_{j1} \\ x_{j2} \\ ... \\ x_{jn} \end{pmatrix}$$

Представим зависимость этих измерений в линейном виде, аналогичном регрессионной модели.

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_m x_m + e\,,\quad y = \hat{y} + e\,,$$

здесь $\hat{y}$ - объяснённая часть измерений, а e - невязка измерений и линейной модели. Подберём неизвестные коэффициенты эмпирической регрессии b из условий

$$\overline{e} = 0\,,\qquad D_e = \frac{1}{n}\sum_{i=1}^{n} e_i^2 \Rightarrow \min_{b_j}\,.$$

Используя то, что наши данные измерений приведены к стандартному масштабу, где $\overline{y} = 0$, $\overline{x}_j = 0$ можно увидеть, что коэффициент $b_0 = 0$. Действительно, вычисляя средние значения и величины дисперсии

$$\overline{y} = b_0 + b_1\overline{x}_1 + b_2\overline{x}_2 + ... + b_m\overline{x}_m + \overline{e}\,,\quad D_y = D_{\hat{y}} + D_e$$

можно заметить, что если $\overline{e} = 0$, то $\overline{y} = \overline{\hat{y}} = 0$ и $b_0 = 0$.

Учитывая это, запишем уравнения эмпирической регрессии в индексном и в матричном виде:

$$y_i = b_1 x_{1i} + b_2 x_{2i} + ... + b_m x_{mi} + e_i\,,\ y_i = \hat{y}_i + e_i\ :$$

$$y = X\cdot b + e\,,\ \ \hat{y} = X\cdot b\,,\ e = y - \hat{y}\,.$$

Запишем также условие минимальности дисперсии невязок в виде:

$$D_e = \frac{1}{n}\sum_{i=1}^{n} e_i^2 = \frac{1}{n}e^T e = \frac{1}{n}(y - X\cdot b)^T\cdot(y - X\cdot b) \Rightarrow \min_b\,,$$

который показывает, что условие минимальности дисперсии невязок эквивалентно главному принципу метода наименьших квадратов (МНК).

$$nD_e = y^T y - y^T Xb - b^T X^T y + b^T X^T X\cdot b \Rightarrow \min_b$$

Учитывая, что минимум положительно определённой квадратичной формы достигается в стационарной точке, получим:

$$n\frac{\partial D_e}{\partial b}_e = -2X^T y + 2X^T X\cdot b = 0 \quad \Rightarrow \quad b = (X^T X)^{-1}\cdot(X^T y)$$

Найденные коэффициенты регрессии b доставляют минимум дисперсии невязки регрессии. Сама невязке вычисляется так:

$$e = y - \hat{y} = y - Xb = y - X\cdot(X^T X)^{-1}\cdot(X^T y)\,.$$

Построенная регрессия с коэффициентами $b = (b_1, b_2, ...b_m)^T$, называемая также линейным трендом, объясняет величину y через величины $x_1, x_2, ...x_m$ не полностью, а лишь частично в силу $e \neq 0$. В качестве меры объяснения удобно

принять величину дисперсии $D_{\hat{y}}$ или, как её называют, величину изменчивости переменной. При этом безразмерный коэффициент $R^2 = D_{\hat{y}} / D_y$ называется коэффициентом детерминации:

$$R^2 = \frac{D_{\hat{y}}}{D_y} = 1 - \frac{D_e}{D_y}.$$

Коэффициент детерминации показывает долю объяснённой дисперсии в общей наблюдаемой дисперсии (изменчивости) объясняемой величины. К свойствам коэффициента необходимо отнести следующее:

$$R^2 \le 1, \qquad \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m} = F_{m,n-m-1}.$$

Последнее соотношение позволяет определить значимость коэффициента детерминации по критерию Фишера.

$$F_{êðèò} = FÐÀÑÏ\ .Î\ ÁÐ.Ï\ Õ(\alpha, m, n-m-1), \quad G_{æ\ ðàñ} = F_{í\ àáë} / F_{êðèò}$$

Коэффициент значим по уровню α, если жёсткость критерия G больше единицы, при этом и сама регрессия называется значимой. Незначимая регрессия не обладает достаточным качеством и не имеет практического применения. Чем выше жёсткость критерия, тем качественнее является регрессия с точки зрения объяснения объясняемой величины.

Числовой пример (часть 3)

В части 1 нашего сквозного примера для наблюдаемых исходных измерений была построена матрица измерений в стандартной форме (центрированная и нормированная). Построим по ней линейную регрессию (тренд), для этого вычислим коэффициенты регрессии b и построим значения тренда $\hat{y}$ и невязок e:

$X =$

x1	x2	x3	x4	y	утренд	e
0.306	-0.454	0.593	-0.756	-0.673	-0.690	0.016
0.066	-0.454	0.091	1.323	0.126	0.330	-0.204
-0.896	-1.079	1.598	-0.756	-0.773	-1.299	0.526
-0.415	1.420	-1.415	-0.756	0.626	0.701	-0.075
0.787	0.795	-0.913	1.323	1.625	1.250	0.375
1.749	-1.079	0.593	-0.756	-0.823	-0.949	0.126
-0.415	-0.454	-0.411	1.323	0.026	0.410	-0.384
-1.377	0.170	0.091	-0.756	-0.703	-0.337	-0.366
1.509	0.170	1.598	-0.756	-0.793	-0.501	-0.293
0.066	-1.079	-0.411	-0.756	-0.763	-0.828	0.064
-1.377	2.045	-1.415	1.323	2.125	1.913	0.212

срзнач	0.000	0.000	0.000	0.000	0.000
дисп	1.000	1.000	1.000	1.000	1.000
ско	1.000	1.000	1.000	1.000	1.000
стандоткл	1.049	1.049	1.049	1.049	1.049

0.000	0.000
0.919	0.081
0.958	0.285
1.005	0.299

Z=X"X=

11.000	-3.729	4.040	-1.954
-3.729	11.000	-7.386	4.014
4.040	-7.386	11.000	-5.503
-1.954	4.014	-5.503	11.000

Zобр=

0.107	0.018	-0.028	-0.002
0.018	0.169	0.103	-0.007
-0.028	0.103	0.199	0.057
-0.002	-0.007	0.057	0.122

X"У=

-3.141
9.054
-8.516
8.113

в=

0.053
0.538
-0.210
0.446

R2=	**0.919**	значим
Fнабл=	16.9069	
Fкрит=	4.53369	
G=	**4.53369**	

5.2. **Теорема Гаусса-Маркова**

Коэффициент детерминации является важным обобщающим критерием качества регрессии, но он не единственный и не отвечает на ряд важных вопросов о соответствии построенной методом наименьших квадратов эмпирической регрессии той теоретической линейной регрессионной связи между случайными величинами Y и X_j.

$$Y = \beta_1 X_1 + \beta_2 X_2 + ... + \beta_m X_m + \varepsilon, \quad \beta_j = const .$$

На поставленный вопрос отвечает теорема Гаусса-Маркова, утверждающая, что **<u>если</u>** специфические остатки ε_n во всех измерениях при наблюдении обладают следующими свойствами (предпосылками МНК [2]):

1. $M(\varepsilon_i) = 0$, несмещ`нность остатков,
2. $D(\varepsilon_i) = \sigma^2 = const$, гомоскедастичность остатков,
3. $\mathrm{cov}(\varepsilon_i, \varepsilon_k) = 0$ при $i \neq k$, некоррелируемость остатков между собой.

Условия 2-3 могут быть записаны в виде $\mathrm{cov}(\varepsilon_i, \varepsilon_k) = I\sigma^2$, $I = \delta_{ik}$.

4. $\mathrm{cov}(\varepsilon_i, X_k) = 0$, некоррелируемость остатков и объясняющих переменных.
5. $\varepsilon_i = N(0, \sigma)$, нормальность остатков.

<u>Тогда</u> построенная по методу МНК эмпирическая регрессия

$$y = b_0 + b_1 x_1 + b_2 x_2 + ... + b_m x_m + e$$

является несмещённой, состоятельной и эффективной оценкой для теоретической регрессии.

Это означает, что оценки $\beta_j^* = b_j$ обладают следующими свойствами:

1. $M(b_j) = \beta_j$, несмещённость

2. $D(b_j) = \sigma^2 (X^T X)^{-1}_{jj} = \frac{\sigma^2}{n} R^{-1}_{jj} \xrightarrow[n\to\infty]{} 0$, состоятельность

3. $D(b_j) = \min\limits_{g(X)}$, эффективность.

Рассмотрим некоторые следствия теоремы Гаусса-Маркова. Величина $\sigma^2 = D(\varepsilon)$ является неизвестной и неизмеримой в наблюдениях, ее эмпирическим аналогом является величина D_e, однако $M(D_e) \neq \sigma^2$, но эту смещённость можно исправить, введя величину

$$s^2 = \frac{n}{n-m-1} D_e = \frac{1}{n-m-1} \sum_{i=1}^{n} e_i ,$$

для которой $M(s^2) = \sigma^2$, то есть смещение отсутствует. Величина s называется стандартной ошибкой регрессии. Через нее выражаются следующие важные величины:

$$s_{b_j} = s\sqrt{(X^T X)^{-1}_{jj}} ,$$

называемые стандартными ошибками коэффициентов регрессии. Кроме того, доказано [1], что отклонение $b_j - \beta_j$ пропорционально величине Стьюдента

$$\frac{b_j - \beta_j}{s_{b_j}} = t_{n-m-1} .$$

Последнее равенство позволяет построить, во-первых, доверительные интервалы для коэффициентов теоретической регрессии с надежностью γ

$b_j - t_\gamma s_{b_j} < \beta_j < b_j + t_\gamma s_{b_j}$, где $t_\gamma = \text{ÑÒÜÞ ÄÎ ÁÐ.2Õ}(1-\gamma, n-m-1)$,

а во-вторых, проверить значимость по уровню α коэффициентов эмпирической регрессии b_j по критерию Стьюдента:

$$t_{j,\text{íàáë}} = \frac{b_j}{s_{b_j}} , \quad G_j = t_{j,\text{íàáë}} / t_{\text{êðèò}} , \quad t_{\text{êðèò}} = \text{ÑÒÜÞ ÄÎ ÁÐ.2Õ}(\alpha, n-m-1) .$$

Коэффициент регрессии b_j является значимым (значимо отличным от нуля), если его жёсткость $G_j > 1$. Значимость коэффициента регрессии говорит о значимости переменной в модели регрессии, что позволяет решать проблему спецификации модели регрессии со стороны отбрасывания незначимых для модели переменных. Однако, отбрасывая незначимую переменную, мы не должны значимо уменьшать коэффициент детерминации. Если это происходит, то незначимая переменная видимо коррелирует с другими уже значимыми переменными. В этом случае отбрасывание переменных нежелательно.

Числовой пример (часть 4**)**

Продолжая сквозной пример, вычислим стандартные ошибки регрессии, доверительные интервалы для коэффициентов теоретической регрессии β_j и значимость эмпирических коэффициентов b_j.

Стандартная ошибка регрессии $s = 0.387$, а ошибки коэффициентов и их доверительные интервалы при заданной надёжности $\gamma = 0.95$ будут:

в=	**0.053**	Sb=	0.126		-0.193		0.299
	0.538		0.159		0.229	$< \beta <$	0.846
	-0.210		0.172		-0.545		0.126
	0.446		0.135		0.184		0.708

Расчёт значимости коэффициентов показывает следующее:

	0.170874		0	R2=	**0.663**	незачим
G=	1.382713	bзнач=	0.538	Fнабл=	2.947689	
	-0.49669		0	Fкрит=	4.533689	
	1.352548		0.446	G=	**0.650175**	

Видим, что коэффициенты b_1 и b_3 незначимы при заданном уровне значимости $\alpha = 0.05$. Отбрасывая незначимые переменные получим регрессию только по двум переменным x_2 и x_4, однако коэффициент детерминации при этом резко уменьшается и становится не значимым. Переменные x_1 и x_3 вместе отбрасывать нельзя, так как они влияют на значимые переменные. Отбросим только переменную x_1, тогда получим следующее:

	0	R2=	**0.951**	зачим
bзнач=	0.538	Fнабл=	29.42684	
	-0.210	Fкрит=	4.533689	
	0.446	G=	**6.490705**	

Получается что, регрессионная модель вида $\hat{y} = b_2 x_2 + b_3 x_3 + b_4 x_4$ без учёта x_1 (урожайность земельного участка) имеет более высокий коэффициент детерминации, чем с её учётом. Так решаются некоторые вопросы спецификации модели.

5.3. **Проверка предпосылок МНК по входным данным**

Для того, чтобы построенная методом МНК линейная регрессия обладала нужными качествами, необходимо выполнение условий (предпосылок) Гаусса-Маркова на наблюдаемые переменные X_j, Y. Но об их свойствах мы можем судить только по их наблюдениям, поэтому в наблюдательных данных не должны проявляться нарушения предпосылок Гаусса-Маркова. К нарушениям этих предпосылок относятся:

- гетероскедастичность данных $(D(e_i) \neq const,\ \operatorname{cov}(e_i, x_k) \neq 0)$,

- автокорреляция данных $\mathrm{cov}(e_i, e_k) \neq 0,$ в частности $\mathrm{cov}(e_i, e_{i-1}) \neq 0$

- мультиколлинеарность данных $(\mathrm{cov}(x_j, x_k) \approx 1)$.

Рассмотрим проявление этих нарушений, их значимость и методы устранения.

Гетероскедастичность (неодинаковость) остатков регрессии, их зависимость от переменных $x_j, y, \hat{y}$. Это явление можно обнаружить, построив графические зависимости для e от указанных величин, или вычислив коэффициенты корреляции $r_{e,x_j}, r_{ey}, r_{e\hat{y}}$, значимость которых говорит о наличии гетероскедастичности в данных.

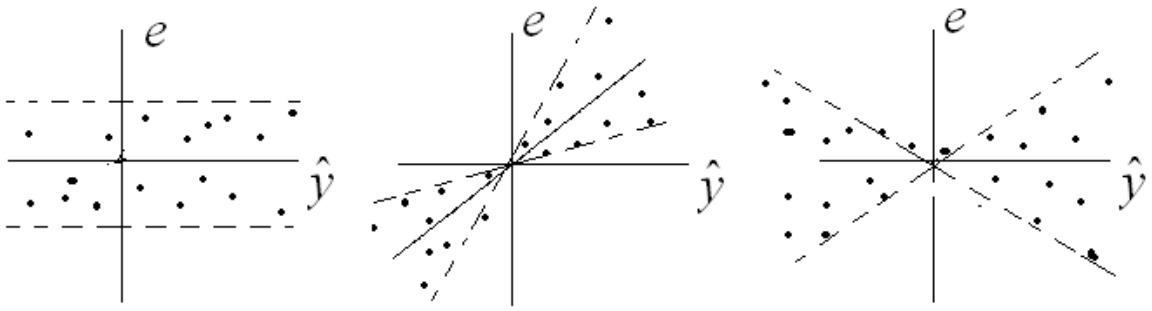


Рис. 5.1. Гетероскедастичность данных. Слева случай отсутствия гетероскедастичности

Согласно тесту Глейзера [9,10], построив парную регрессионную зависимость $e = r_{e,x_j} x_j + w$, можно проверить её значимость по критерию Стьюдента, но при этом гетероскедастичность может присутствовать в данных и в случае незначимого коэффициента детерминации. Наиболее популярным и строгим является тест ранговой корреляции Спирмена, согласно которому вычисляется коэффициент ранговой корреляции:

$$Sr_{e\hat{y}} = 1 - \frac{\sum_{i=1}^{n} [rang(e_i^2) - rang(\hat{y}_i^2)]^2}{(n-1)n(n+1)/6},$$

где $rang(e_i^2) = \text{ÐÀÍ}\ \tilde{A}(å_i^2, e^2, 1)$ есть порядковый номер элемента e_i^2 в массиве квадратов остатков e^2, расположенном по возрастанию. Если ранги остатков и объяснённой части $\hat{y}$ точно соответствуют друг другу (e зависит от $\hat{y}$), то $Sr_{e,\hat{y}} = 1$, если же ранги распределены случайно друг относительно друга, то $Sr_{e,\hat{y}} \approx 0$. Доказано [1], что:

$$\frac{Sr_{e\hat{y}}\sqrt{n-2}}{\sqrt{1-Sr_{e\hat{y}}^2}} = t_{n-2}$$

и, следовательно, для определения значимости $Sr_{e,\hat{y}}$ можно воспользоваться критерием Стьюдента при заданном уровне значимости α. Если коэффициент ранговой корреляции оказывается значимым, то в наблюдательных данных присутствует гетероскедастичность, что может привести к смещению построенной эмпирической регрессии по отношению к истинной теоретической регрессии между наблюдаемыми переменными. В случае гетероскедастичности необходимо

отказаться от измеренных данных и повторить их заново, или провести их преобразование (видоизменение, отсев).

Автокорреляция (самозависимость) остатков регрессии является простейшим частным случаем зависимости остатков между собой. При автокорреляции 1-го порядка соседние остатки регрессии жёстко связаны между собой.

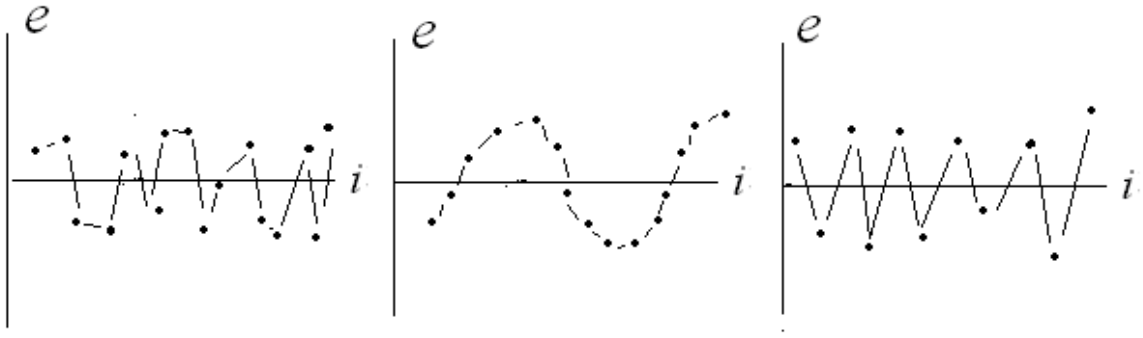


Рис. 5.2. Автокорреляция в данных, $r_{e1}=0$, положительная автокорреляция $r_{e1}>0$, отрицательная автокорреляция $r_{e1}<0$.

Моделью автокорреляции 1-го порядка является регрессионная зависимость остатков со сдвигом на один элемент $e_i = r_{e1}e_{i-1} + w_i$. Если обычный коэффициент корреляции r_{e1}, значим, то это говорит о наличии автокорреляции, однако при его незначимости автокорреляция в данных всеже может быть. Для более строгого определения автокорреляции обычно используется критерий Дарбина-Уотсона DW_{nm} [9,10]:

$$r_{e1} = \frac{\sum_{i=2}^{n} e_i e_{i-1}}{\sum_{i=1}^{n} e_i^2}, \qquad DW_{nm} = \frac{\sum_{i=2}^{n} (e_i - e_{i-1})^2}{\sum_{i=1}^{n} e_i^2} = \frac{\sum_{i=2}^{n} e_i^2 - e_i e_{i-1} + e_{i-1}^2}{\sum_{i=1}^{n} e_i^2} \approx 2(1 - r_{e1}).$$

Распределение критерия DW отличается крутизной наклона от Стьюдентовского (рис. 2.8) и приводится на рис.5.3, где показаны критические области при проверке гипотезы $H_0 = \{\rho_{ä1} = 0\}$:

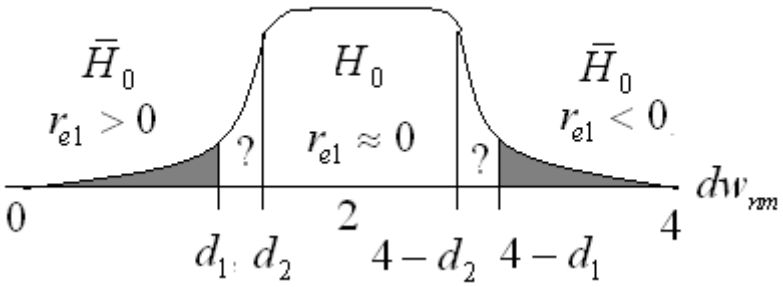


Рис. 5.3. Критерий Дарбина=Уотсона

Значение критерия Дарбина-Уотсона $0 \le DW \le 4_{nm}$, что соответствует значениям коэффициента корреляции $-1 \le r_{e1} \le +1$. Критические точки распределения определяются по таблицам в приложение №2 (в Exel нет обратной функции этого

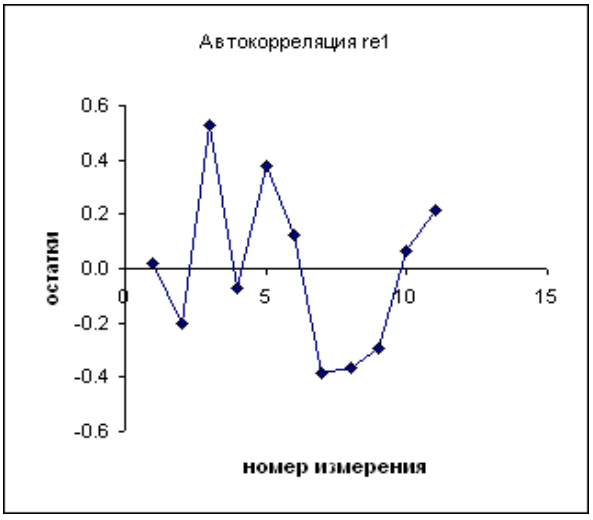
распределения). В силу особенностей распределения критерия они как бы расщепляются на две пары точек d_1, d_2, и $4-d_1, 4-d_2$ между которыми критерий не даёт ответа на вопрос о значимости коэффициента автокорреляции, в остальном всё аналогично критерию Стьюдента.

Мультиколинеарность («параллельность») измеряемых величин говорит о наличии среди измеряемых величин жёстко коррелирующих пар величин. Такая зависимость измеряемых в наблюдениях величин приводит к большим ошибкам коэффициентов регрессии или даже невозможности построить саму регрессию в силу необратимости матрицы парных корреляций. Именно по этой матрице может быть определена жёсткая связь переменных. Одну из коррелирующих величин необходимо исключить из спецификации регрессионной модели. Однако, необходимо помнить, что исключаемая переменная может опосредованно влиять или зависеть от других переменных модели. Поэтому, для исключения переменной из модели, нужен анализ не только парных корреляций, но и частных (очищенных) корреляций, или преобразовать эту переменную каким-либо образом. Подобные преобразования рассматриваются ниже.

Числовой пример (часть 5)

Рассмотрим выполнение предпосылок Гаусса-Маркова в нашем сквозном примере статистического анализа.

Наличие автокорреляции можно допустить, анализируя график остатков от номеров измерений, по крайней мере, в первой и второй половине измерений. Однако в целом, вычисленный коэффициент корреляции остатков $r_{e1} = 0.07$ мал, а критерий Дарбина-Уотсогна наблюдается в районе значения 2, казалось бы, что автокорреляция 1-го порядка отсутствует. Но критические значения критерия d_1, d_2 при нашем объёме выборки таковы, что наблюдаемое значение критерия находится между ними. Вывод состоит в том, что критерий не может определить наличие или отсутствие автокорреляции.



Тест Дарбина-Уотсона

$r_{e1} =$ 0.070

DWнабл=	1.859
d1крит=	0.440
d2крит=	2.283

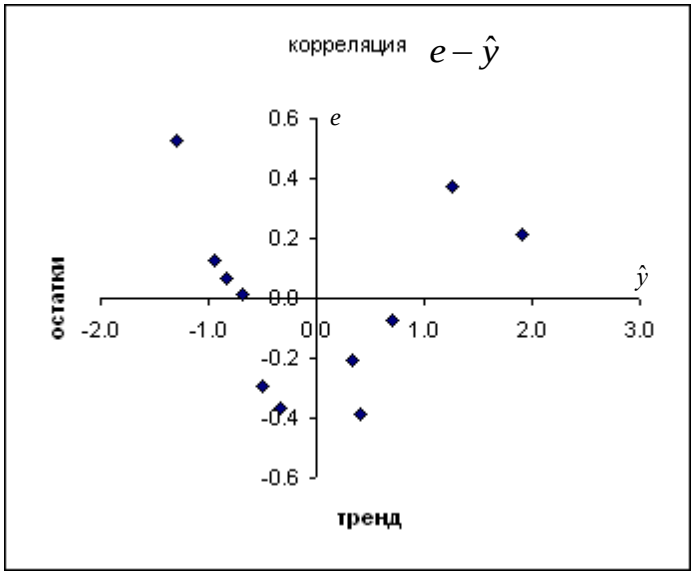
d1крит< DWнабл<d2крит

Наличие или отсутствие автокорреляции
тест не определяет.

Требуются более серьезные исследования, но отметим, что при объёме входных данных $n \geq 19$ в нашем примере ответ о наличии автокорреляции был бы отрицательным.

Исследование гетероскедастичности в нашем примере будем проводить, анализируя связь $e - \hat{y}$, так как объяснённая часть $\hat{y}$ наблюдений является линейной комбинацией наблюдаемых величин. Как графически (шарообразность изображения остатков) на плоскости, так и аналитически по тестам Глейзера и Спирмена видим, что значимое наличие гетероскедастичности данных не устанавливается.

$\hat{y}$	e
-0.690	0.016
0.330	-0.204
-1.299	0.526
0.701	-0.075
1.250	0.375
-0.949	0.126
0.410	-0.384
-0.337	-0.366
-0.501	-0.293
-0.828	0.064
1.913	0.212



РАНГ $\hat{y}^2$	РАНГ e^2
5	1
1	5
10	11
6	3
9	9
8	4
3	10
2	8
4	7
7	2
11	6

Тест Глейзера

$r_{e\hat{y}} =$ 0.000

tнабл= 0

tкрит= 2.262

незначим

Тест Спирмена

$r_{e1} =$ 0.082

tнабл= 0.246

tкрит= 2.262

незначим

Гетероскедастичности в данных нет

5.4. О нелинейной регрессии и инструментальных переменных

Иногда линейная модель регрессии бывает недостаточной, с точки зрения её качества и значимости, тогда могут быть использованы нелинейные модели. В простейшей форме нелинейность может быть учтена путем введения инструментальных переменных $z_i=\varphi_i(x_i)$, которые входят в модель регрессии обычным линейным образом. Например, нелинейная модель 2-го порядка может быть построена следующим образом (рис. 5.4):

$$\hat{y}(x) = b_0 + b_1 x + b_2 x^2 + e \quad \Rightarrow \quad \hat{y}(x, z_2) = b_0 + b_1 x + b_2 z_2 + e,$$

где $z_2 = x^2$ - инструментальная переменная.

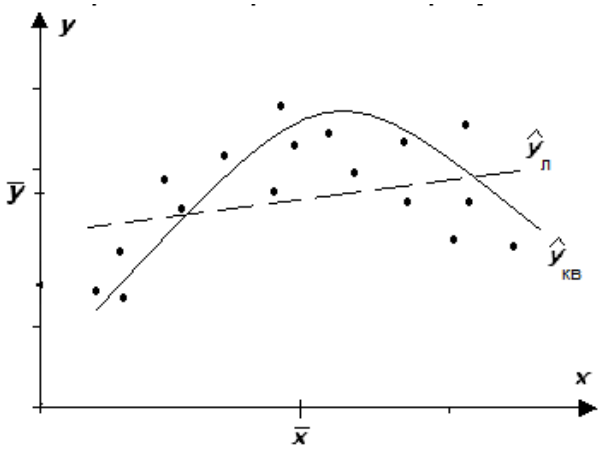


Рис.5.4 Кривая нелинейной среднеквадратической регрессии 2-го порядка

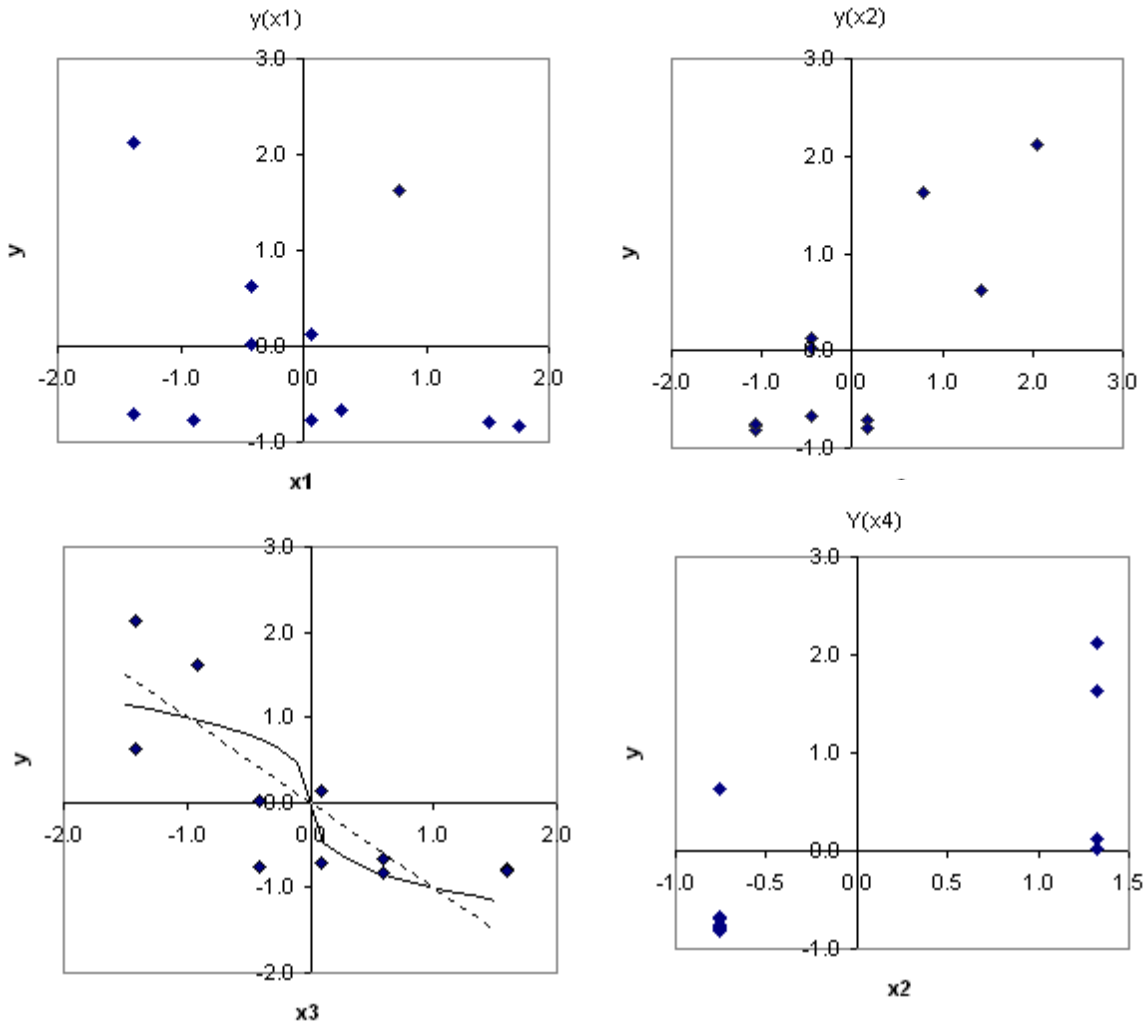
Часто в качестве инструментальных переменных используются степенная функция $z = x^{\delta}$, логарифмическая $z = \ln(\delta + x)$, показательная $z = e^{\delta x}$ и иногда тригонометрическая $z_r = Sin(\omega x + \delta)$ для выявления циклических факторов в зависимостях. Введение новых членов в модель регрессии, в том числе и инструментальных, оправдано тогда, когда возрастает качество регрессии. Например, значимо повышается коэффициент детерминации, ликвидируется мультиколлинеарность переменных, и т.д.

Таким образом, замена или введение новых переменных в регрессионную модель - это задача исследователя в конкретной предметной области деятельности.

Числовой пример (часть 6)

Опишем введение нелинейности в модель через инструментальные переменные в нашем сквозном примере. Для этого, построим графически корреляционное поле в плоскостях (y, x_j) и проанализируем расположение на них точек измерений. Можно заметить по тесноте точек, что зависимость измерений объясняемой переменной y от переменной x_1 очень слаба в отличие от других трёх переменных. Кроме того, видно, что расположение точек на плоскости (y, x_3) группируется не

вдоль прямой линии, как на плоскости (y, x_2), а вдоль некоторой кривой, похожей на кубическую параболу.



Поэтому ввёдем вместо переменной x_3 новую инструментальную переменную $z_3 = \sqrt[3]{x_3}$ и, тем самым, построим регрессионную модель вида:

$$\acute{o} = b_1\tilde{o}_1 + b_2x_2 + b_3\sqrt[3]{x_3} + b_4x_4 + e\,.$$

Эта модель уже нелинейная, но строится обычным образом после преобразования входных данных, при этом коэффициент детерминации повышается от 0.919 до 0.928. В нашем примере это повышение незначительное, но подобные преобразования могут привести и к серьёзному улучшению качества регрессионной модели.

x1	x2	x3	x4	y	у-тренд **линейн**	е	у-тренд нелин	е
0.306	-0.454	0.593	-0.756	-0.673	-0.690	0.016	-0.723	0.049
0.066	-0.454	0.091	1.323	0.126	0.330	-0.204	0.313	-0.186
-0.896	-1.079	1.598	-0.756	-0.773	-1.299	0.526	-1.276	0.503
-0.415	1.420	-1.415	-0.756	0.626	0.701	-0.075	0.766	-0.140
0.787	0.795	-0.913	1.323	1.625	1.250	0.375	1.242	0.384
1.749	-1.079	0.593	-0.756	-0.823	-0.949	0.126	-0.922	0.099
-0.415	-0.454	-0.411	1.323	0.026	0.410	-0.384	0.420	-0.393
-1.377	0.170	0.091	-0.756	-0.703	-0.337	-0.366	-0.423	-0.280
1.509	0.170	1.598	-0.756	-0.793	-0.501	-0.293	-0.552	-0.241
0.066	-1.079	-0.411	-0.756	-0.763	-0.828	0.064	-0.773	0.010
-1.377	2.045	-1.415	1.323	2.125	1.913	0.212	1.929	0.196

$$\tilde{o} = b_1\tilde{o}_1 + b_2x_2 + b_3\sqrt[3]{x_3} + b_4x_4$$

Линейная рег-рес.

R2= **0.919**

S= **0.387**

Нелинейная регрес.

R2= **0.928**

S= **0.364**

6. Задачи факторного анализа

В предыдущих главах рассматривались статистические данные, полученные в наблюдениях за величинами характеристик объектов при помощи того или иного измерительного прибора-инструмента. Такие данные объединялись в матрицу измерения *X*, а наблюдаемые величины называли измеримыми факторами. Однако в наблюдениях часто замечено, что при анализе данных имеются некоторые непосредственно не измеримые, но весьма важные факторы, объясняющие наблюдаемые явления. Эти скрытые (**латентные**) факторы являются или причиной наблюдаемых измерений, или их обобщением. Например, такие понятия, как надёжность изделия или конструкции, её экономичность, экологичность или эргономичность неизмеримы непосредственно, но они связаны с измерениями величин предельно допустимых нагрузок, вибро-шумовых величин, химического состава, величин калибровочных настроек приборов и многого другого. Другой пример следующий, студенческая зачетная книжка есть матрица измерений его знаний, но она есть отражение других, скрытых факторов, таких как способность, усидчивость, мотивируемость, здоровье, склонности к точным, естественным или гуманитарно-художественным видам деятельности.

Возникает вопрос, как методами статистического анализа данных измерений выделить и вычислить эти скрытые общие факторы, а также объяснить ими наблюдаемые измеримые факторы. Кроме того, желательно иметь весьма ограниченный набор таких общих факторов, который бы объяснял наблюдения с достаточно хорошей точностью.

Мы будем строить факторную модель следующего вида [1,2]:

$$x_j = \hat{O}_j(f_1, f_2 .... f_p) + \varepsilon_j .$$

Здесь $x_j = (x_{1j}, x_{2j} .... x_{nj})^{\grave{O}}$ вектора измеримых *j*-факторов, $f = (f_1, f_2 .... f_p)^{\grave{O}}$- вектор общих (латентных) факторов $(p < m)$, $\varepsilon = (\varepsilon_1, \varepsilon_2 .... \varepsilon_m)^{\grave{O}}$ - вектор частных (специфических) факторов, присущих именно измеримым факторам.

6.1. Линейная факторная модель

Наиболее простой и популярной факторной моделью является линейная факторная модель

$$x_j = \lambda_{j1} f_1 + \lambda_{j2} f_2 + ... + \lambda_{jp} f_p + \varepsilon_j .$$

Каждый измеримый *j*-фактор представляется в модели суммой линейной общефакторной части $\hat{x}_j = \lambda_{j1} f_1 + \lambda_{j2} f_2 + ... + \lambda_{jp} f_p$ и характерной специфической части ε_j.

В матричной форме модель по всем измерениям измеримых факторов выглядит следующим образом:

$$X = \Lambda F + \varepsilon,$$

где $X = (x_{ij})$ матрица измерений $(n \times m)$, $\Lambda = (\lambda_{ik})$ - матрица факторных нагрузок $(n \times p)$, $F = (f_{kj})$ - матрица значений общих факторов $(p \times m)$, $\varepsilon = (\varepsilon_{ij})$ -матрица специфичности $(n \times m)$.

$$X = \begin{pmatrix} x_{11} & x_{21} & \dots & x_{m1} \\ x_{12} & x_{22} & \dots & x_{m2} \\ \dots & \dots & \dots & \dots \\ x_{1n} & x_{2n} & \dots & x_{mn} \end{pmatrix}, \Lambda = \begin{pmatrix} \lambda_{11} & \lambda_{21} & \dots & \lambda_{p1} \\ \lambda_{12} & \lambda_{22} & \dots & \lambda_{p2} \\ \dots & \dots & \dots & \dots \\ \lambda_{1n} & \lambda_{2n} & \dots & \lambda_{pn} \end{pmatrix},$$

$$F = \begin{pmatrix} f_{11} & f_{21} & \dots & f_{n1} \\ f_{12} & f_{22} & \dots & f_{n2} \\ \dots & \dots & \dots & \dots \\ f_{1p} & f_{2p} & \dots & f_{np} \end{pmatrix}, \varepsilon = \begin{pmatrix} \varepsilon_{11} & \varepsilon_{21} & \dots & \varepsilon_{m1} \\ \varepsilon_{12} & \varepsilon_{22} & \dots & \varepsilon_{m2} \\ \dots & \dots & \dots & \dots \\ \varepsilon_{1n} & \varepsilon_{2n} & \dots & \varepsilon_{mn} \end{pmatrix}.$$

Поскольку через вводимые общие факторы мы хотим линейно объяснять все измеримые факторы, то будем требовать от этих факторов (вернее их значений) выполнения условий теоремы Гаусса-Маркова, а именно:

1. $M(\varepsilon_k) = 0$ несмещённость специфических факторов,

2. $\mathrm{cov}(\varepsilon_k, \varepsilon_r) = 0$ некоррелируемость специфических факторов,

3. $\mathrm{cov}(\varepsilon_k, \varepsilon_k) = \sigma_k^2 = const$ гомоскедастичность специфических факторов,

4. $\varepsilon_k = N(0, \sigma_k)$ нормальность специфических факторов.

5. $\mathrm{cov}(f_k, \varepsilon_r) = 0$ некоррелируемость общих и специфических факторов между собой.

Кроме того, будем требовать от модели также некоррелируемость общих факторов $\mathrm{cov}(f_k, f_r) = 0$ и упорядоченность их факторной дисперсии $\mathrm{cov}(f_k, f_k) = D_k$ по убыванию $D_1 > D_2 > \dots > D_p$. Последнее требование позволяет назвать первый фактор главным, т.к. именно он определяет наибольшую долю факторную изменчивость $D_1 / \sum_k D_k$. Чем больше первые факторные дисперсии и чем меньше последние, тем выше качество факторной модели. Количество общих факторов p должно быть не большим, по крайней мере $p < m$, это позволит провести уменьшение размерности задачи анализа измеренных величин. Такое сжатие информации об измеренных объектах позволит эффективно и наглядно решать задачи о классификации наблюдаемых объектов в задачах кластерного анализа.

Как же ввести общие факторы в наблюдаемый нами массив измерений? Для понимания главной идеи факторного анализа вернёмся к задаче парного регрессионного анализа. Построим для корреляционного поля $(x_i; y_i)$ в центральной нормированной системе координат новую систему координат $(f_1; f_2)$, повёрнутую на угол θ и в каждой из них прямоугольник, охватывающий корреляционное поле. На рисунке 6.1 видно, что размеры этих прямоугольников различны, но ведь они соответствуют дисперсиям наблюдаемых величин, причём как видим из рисунка $\sigma_x > \sigma_y; \sigma_{f_1} > \sigma_{f_2}$, $\sigma_{f_1} > \sigma_x; \sigma_{f_2} < \sigma_y$.

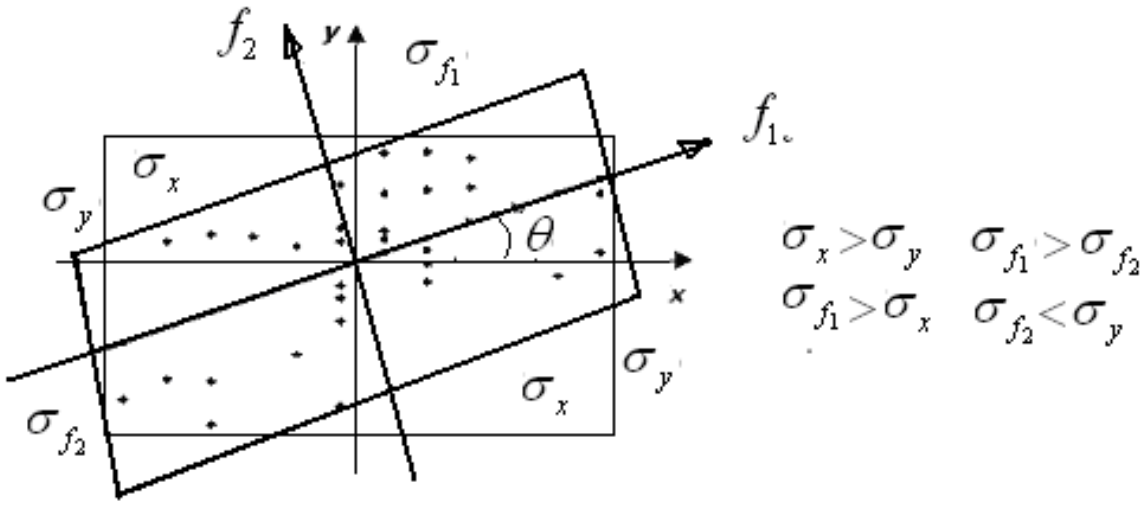


Рис. 6.1. Идея построения новых факторов (главных координат) путем вращения измеримых факторов.

В новой системе координат $(f_1; f_2)$ первая дисперсия увеличилась, а вторая уменьшилась. Это подсказывает, что новые факторы можно вводить как новую систему координат, повёрнутую относительно исходной системы, где произведены измерения. Преобразование поворота прямоугольной системы координат известно:

$$\begin{aligned} f_1 &= x\cos\theta - y\sin\theta \\ f_2 &= x\sin\theta + y\cos\theta \end{aligned} \quad \text{èëè} \quad \begin{pmatrix} f_1 \\ f_2 \end{pmatrix} = \begin{pmatrix} \cos\theta & -\sin\theta \\ \sin\theta & \cos\theta \end{pmatrix} \cdot \begin{pmatrix} x \\ y \end{pmatrix}$$

Угол θ в матрице поворота T можно выбрать из различных соображений, например, в методе центройд [3], ось главного фактора f_1 направляется на наиболее удалённую точку корреляционного поля, а ось фактора f_2 должна быть перпендикулярной к оси f_1. Можно направить первую ось вдоль линии регрессии, при этом $tg\theta = r_{xy}$, так как регрессия имеет вид $\hat{y} = r_{xy}x$.

Наиболее известно построение многомерной линейной факторной модели методом главных координат.

6.2. Метод главных координат

В основе метода главных координат лежит линейное преобразование поворота измеренных величин x_j в новые координаты f_k, которые и будут общими факторами, причём размерность пространства при повороте не меняется $p = m$. Запишем преобразование через матрицу поворота T

$f = T \cdot x$; $x = A \cdot f$, где $A = T^{-1}$ обратная матрица поворота.

Матрица поворота должна обладать рядом свойств, обеспечивающих выполнение требований Гаусса-Маркова к линейной факторной модели. Для этого матрица поворота должна быть [9]:

- ортогональной $T^{-1} = T^T = A$ для обеспечения ортогональности новой системы координат;

- преобразование вращения должно диагонализировать корреляционную матрицу в новой системе координат.

Рассмотрим преобразование корреляционной матрицы:

пусть $R = \frac{1}{n} X^T \cdot X$ симметричная корреляционная матрица в нормированной исходной системе измерений $R^T = R$, а $F = T \cdot X$ преобразованная вращением матрица измерений в новых факторных переменных, она несимметрична $F^T = X^T \cdot T^T$. Тогда корреляционная матрица в новых координатах будет следущей:

$$Q^T = \frac{1}{n} F \cdot F^T = \frac{1}{n} TX \cdot X^T T^T = TRT^T .$$

Построим матрицу поворота так, чтобы корреляционная матрица $Q = Q^T$ была бы диагональной, что обеспечит независимость новых факторных переменных. Это требование будет выполнено, если матрица поворота построена из собственных векторов корреляционной матрицы R. Собственные вектора $l \neq 0$ находятся из условия, $R \cdot l = \lambda \cdot l$ или $(R - \lambda E) \cdot l = 0$ которое разрешимо при значении величин λ, удовлетворяющих характеристическому уравнению $|R - \lambda E| = 0$. Это алгебраическое уравнение порядка m для симметричной положительно определенной матрицы R имеет m вещественных положительных корней λ_k, называемых собственными числами корреляционной матрицы. Упорядочив их по убыванию значения и найдя каждому из них соответствующий собственный вектор l_k единичной длины, можем построить следующую матрицу поворота $T = (l_1, l_2, ...l_3)$. Корреляционная матрица в факторных переменных будет диагональной $Q = TRT^T = diagan(\lambda_j)$ с диагональными элементами $\lambda_1 > \lambda_2 > ... > \lambda_m > 0$. Такое преобразование матрицы оставляет инвариантным её след

$Sp[R] = Sp[Q]$, а так как $r_{jj} = D_j = 1$ и $D_{f_j} = \lambda_j$, то выполнено следующее $D_{f_1} + D_{f_2} + ... + D_{f_m} = D_{x_1} + D_{x_2} + ... + D_{x_m} = m$.

6.3. Факторизация модели главных координат

В модели главных координат нет никакой потери информации и точности в измерениях, просто осуществлён переход к новой системе упорядоченных факторных переменных. Проведём факторизацию модели путём выбора общих и специфических факторных переменных, для чего выразим через них измеренные величины: $x = T^{-1} f$, $T^{-1} = A$,

$$x_j = [a_{j1} f_1 + a_{j2} f_2 + ... + a_{jp} f_p] + \{a_{jp+1} f_{p+1} ... + a_{jm} f_m\}.$$

Примем первые p факторных переменных с наибольшими факторными дисперсиями за общие факторы, которые определяют общефакторную часть $\hat{x}_j$ измерений, а остальные отнесём к специфической части ε_j измеримых величин. При этом для общих и специфических факторов в силу их ортогональности выполняются все условия Гаусса – Маркова.

Поскольку теперь $p < m$, то доля факторизации определяется как $\hat{O}(p) = \frac{1}{m} \sum_{k=1}^{p} D_{f_k} < 1$. Задаваясь долей факторизации $\hat{O}_0$, можно определить количество необходимых общих факторов p из условия $\hat{O}(p) \geq \hat{O}_0$.

Большой проблемой проведённой факторизации является смысловая интерпретация полученных общих факторов [10], поскольку они являются линейной комбинацией абсолютно всех измеряемых величин. Для разрежения матрицы факторных нагрузок (получения нулевых или незначимых элементов) иногда проводится дополнительное вращение или даже переход от ортогональных координат к косоугольным координатам. В итоге каждый общий фактор связывается только с частью измеримых величин, что упрощает его интерпретацию. Обычной интерпретацией общего фактора является некий обобщённый уровень значения, например, уровень развития, уровень потребления, уровень качества, уровень образования и т.д.

Числовой пример (часть 7)

Проведём для рассматриваемых в примере измерений выделение и интерпретацию главных факторов, объясняющих не менее 75% изменчивости наблюдаемых величин. Согласно методу главных координат провёдем преобразование измеренных величин путём ортогонального поворота. Матрица поворота T определяется через собственные вектора корреляционной матрицы R. Однако, к сожалению в Excel нет функции определения собственных чисел и векторов матрицы, поэтому можно воспользоваться другими приложениями, например, в сис-

теме MatLab оператор *[T,Q]=equ(R)* формирует матрицу поворота T и преобразованную к главным координатам корреляционную матрицу Q. Преобразованная матрица Q будет диагональной, на диагонали которой находятся убывающие по величине положительные собственные числа матрицы R и являющиеся дисперсиями главных координат.

$T =$

-0.385	0.814	-0.431	0.058
0.551	-0.007	-0.585	-0.596
-0.588	-0.102	0.231	-0.769
0.450	0.572	0.648	-0.226

$T^{-1} = T^T =$

-0.385	0.551	-0.588	0.450
0.814	-0.006	-0.102	0.572
-0.431	-0.585	0.231	0.648
0.058	-0.596	-0.769	-0.226

$R =$

1.000	-0.339	0.367	-0.178
-0.339	1.000	-0.671	0.365
0.367	-0.671	1.000	-0.500
-0.178	0.365	-0.500	1.000

$Q = T^T RT =$

2.252	0.000	0.000	0.000
0.000	0.832	0.000	0.000
0.000	0.000	0.611	0.000
0.000	0.000	0.000	0.305

$\Sigma =$ 4.000

25.0%	25.0%	25.0%	25.0%
25.0%	50.0%	75.0%	100.0%

$\Sigma =$ 4.000

56.3%	20.8%	15.3%	7.6%
56.3%	77.1%	92.4%	100.0%

Суммарная дисперсия (изменчивость) всех наблюдаемых величин в стандартном масштабе равна количеству переменных $m = 4$ и сохраняется при повороте (инвариантность следа матрицы). Распределение изменчивости по координатам сильно изменилось, видим, что первый главный фактор f_1 объясняет уже 56.3% изменчивости, а для объяснения 75% достаточно первых двух факторов f_1, f_2. Поясним смысл главных факторов, т.к. $f = Tx$, то:

$$f_1 = -0.39x_1 + 0.81x_2 - 0.43x_3 - 0.06x_4,$$

$$f_2 = 0.55x_1 - 0.007x_2 - 0.59x_3 - 0.6x_4.$$

Факторные нагрузки разнонаправлены по знаку и значительны по величине.

Рассмотрим для сравнения объяснение величины y через главные факторные переменные и измеренные переменные. Пересчитаем матрицу измерения X в главные координаты $F = X \cdot T$ и построим на них линейную регрессию переменной y.

$$\hat{y}_{mf} = b_{1f} \cdot f_1 + b_{2f} \cdot f_2 + b_{3f} \cdot f_3 + b_{4f} \cdot f_4$$

$F = X \cdot T =$

f1	f2	f3	f4
-1.057	-0.241	-0.219	0.003
0.266	0.804	1.115	-0.094
-1.528	-1.317	0.896	-0.466
1.434	-0.636	-1.467	0.388
1.267	1.485	-0.158	-0.025
-1.957	0.938	-0.476	0.459
0.747	0.463	1.207	0.264
0.231	-1.564	0.026	-0.081
-1.767	0.632	-0.871	-1.072
-0.718	-0.330	0.018	1.133
3.084	-0.234	-0.071	-0.509

$F^T \cdot F =$

24.775	0.001	0.001	0.001
0.001	9.152	0.000	0.000
0.001	0.000	6.717	0.000
0.001	0.000	0.000	3.355

$(F^T \cdot F)^{-1} =$

0.040	0.000	0.000	0.000
0.000	0.109	0.000	0.000
0.000	0.000	0.149	0.000
0.000	0.000	0.000	0.298

0.000	0.000	0.000	0.000		
2.252	0.832	0.611	0.305	$F^T \cdot y =$	14.855 2.892 -0.648 -0.861
1.501	0.912	0.781	0.552		
1.574	0.957	0.820	0.579		

$b =$

0.600
0.316
-0.096
-0.257

В построенной регрессии можно отбрасывать последние факторы (столбцы в F) в любом количестве, вплоть до оставления одного единственного главного фактора.

Из приводимых ниже расчетов можно видеть, что при 4-х факторной регрессии её качественные параметры, такие как коэффициент детерминации R^2 и стандартная ошибка S регрессии совпадают с параметрами регрессии, рассмотренной в части 3 нашего примера. При уменьшении количества используемых факторов параметры регрессии ухудшаются, но незначительно. Так регрессия только по единственному главному фактору $\hat{y}_{1f} = b_{1f} \cdot f_1$ имеет коэффициент детерминации $R^2 = 0.81$, что всего на 10% ниже исходного и остаётся значимым по заданному уровню значимости α. При этом снижается размерность задачи анализа данных, сжимается объём данных путём отбрасывания второстепенных данных и, наконец, имеется возможность визуализации корреляционных полей и линейной регрессии в них.

y	$\hat{y}_{4f}$	e	$\hat{y}_{3f}$	e	$\hat{y}_{2f}$	e	$\hat{y}_{1f}$	e
-0.673	-0.690	0.016	-0.689	0.016	-0.710	0.037	-0.634	-0.039
0.126	0.330	-0.204	0.306	-0.180	0.413	-0.287	0.160	-0.033
-0.773	-1.299	0.526	-1.419	0.646	-1.333	0.559	-0.916	0.143
0.626	0.701	-0.075	0.800	-0.174	0.659	-0.033	0.860	-0.234
1.625	1.250	0.375	1.244	0.381	1.229	0.397	0.759	0.866
-0.823	-0.949	0.126	-0.831	0.008	-0.877	0.054	-1.174	0.350
0.026	0.410	-0.384	0.478	-0.451	0.594	-0.568	0.448	-0.421
-0.703	-0.337	-0.366	-0.358	-0.345	-0.356	-0.347	0.138	-0.842
-0.793	-0.501	-0.293	-0.776	-0.017	-0.860	0.067	-1.059	0.266
-0.763	-0.828	0.064	-0.537	-0.226	-0.535	-0.228	-0.431	-0.333
2.125	1.913	0.212	1.782	0.343	1.775	0.350	1.849	0.276
0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
1.000	0.919	0.081	0.898	0.102	0.893	0.107	0.810	0.190
1.000	0.958	0.285	0.948	0.319	0.945	0.328	0.900	0.436
1.049	1.005	0.299	0.994	0.334	0.991	0.344	0.944	0.458
R2=	**0.919**		**0.898**		**0.893**		**0.810**	
S=	**0.387**		**0.432**		**0.443**		**0.591**	
Gf=	3.729		2.927		2.754		1.408	

На рис. 6.1 изображена проекция корреляционного поля на плоскость $f_1 - f_2$, на которой, как мы знаем, можно увидеть 77% изменчивости измеренных наблюдений. Это позволит классифицировать наблюдаемые объекты на классы, группируя объекты по величине их взаимного расстояния или расположению в квадрантах плоскости.

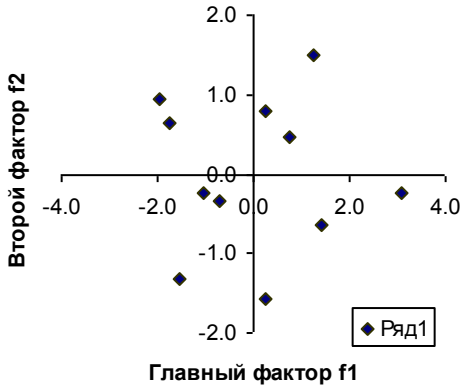


Рис. 6.2. Изображение корреляционного поля наблюдаемых объектов на плоскости первых двух главных факторов.

Задачу классификации объектов можно решать и на прямой единственного главного фактора f_1. На рис. 6.3 представлено не только корреляционное поле объектов, но линия тренда для объясняемой величины y.

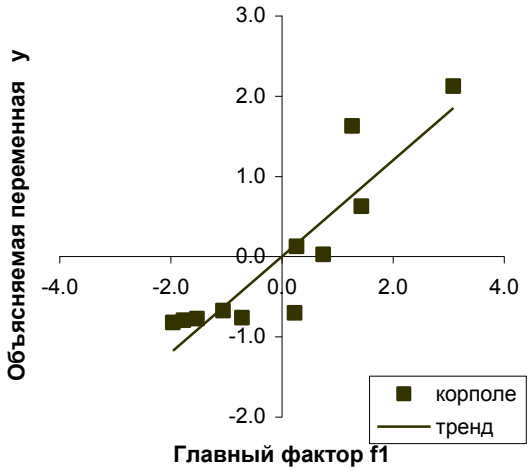
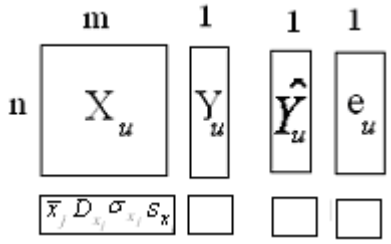


Рис. 6.3. Изображение корреляционного поля наблюдаемых объектов на прямой главного фактора и тренда объясняемой переменной y.

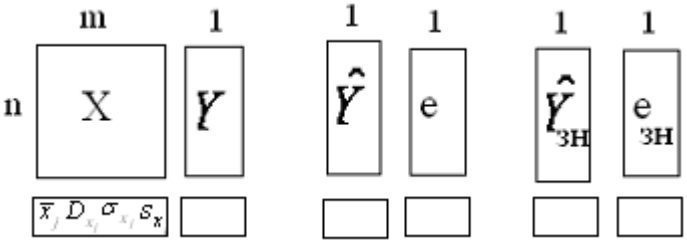
Задания для выполнения расчётно-графических работ.

1. Выбрать входные данные факторов X и У из предложенных ниже по номеру варианта **Nвар**, соответствующему номеру по списку преподавателя и номера Вашей группы **Nгр** и записать их в виде матрацы измерений.



Объём многофакторной выборки X,У вычисляется как **n=10+ номер вариан**та , а величина **N=OCTAT[Номер группы/10**] и равна последней цифре в номере Вашей группы.
Рекомендуется и поощряется использование в качестве входных данных статистические данные, полученные или используемые Вами в определенной предметной деятельности.

2. Перейти к стандартной форме статистических данных X,Y, вычислив средние значения факторов $\overline{o}_j$, дисперсию D_{x_j}, СКО= σ_{x_j} и стандартные отклонения s_{x_j}.

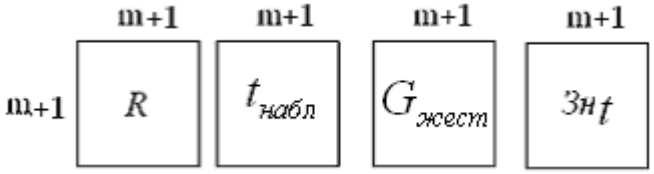


$R_0^2 =$ $R_{çí}^2 =$

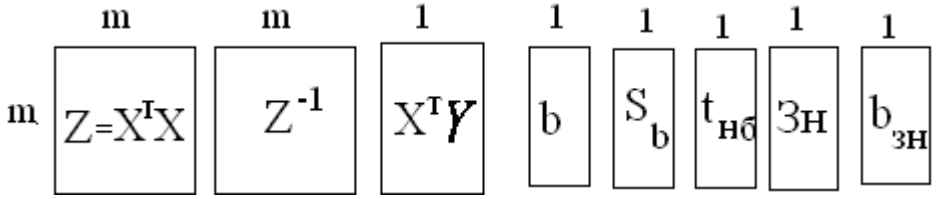
3. Построить матрицу парных корреляций $R=\left(r_{ij}\right)$ измеряемых величин x_i и установить наличие значимых по заданному уровню корреляции α согласно критерию Стьюдента

$$\left| \frac{r_{ij}\sqrt{n-m-1}}{\sqrt{1-r_{ij}^2}} \right| \le t_{кр} \qquad G_{æ\ åñò} = \frac{t_{í\ àáë}}{t_{êðèò}}\ .$$

Имеются ли среди корреляций У и X незначимые переменные x_i?

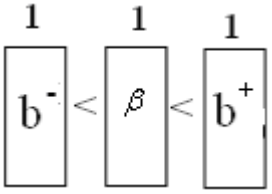


4. Построить коэффициенты линейной регрессии b, ошибки регрессии e, тренд $\hat{Y}$ и построить точечные графики на плоскостях $(e,\hat{Y})$, (e,i).

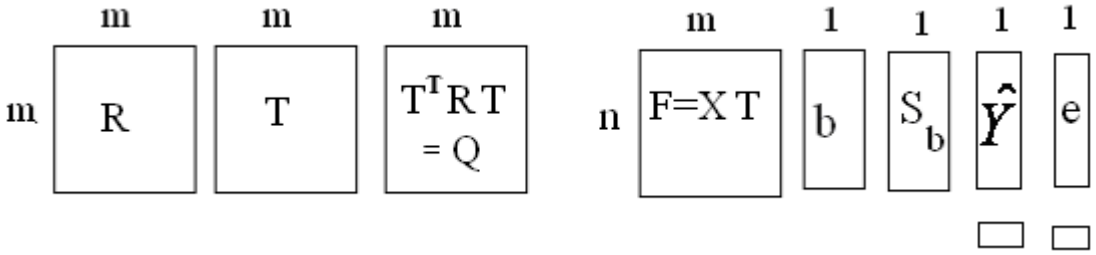


Пересчитать тренд $\hat{Y}$ из стандартной формы в реальные масштабы.

5. Вычислить коэффициент детерминации R_0^2 и установить его значимость, соответствующую заданному уровню значимости альфа. Вычислить стандартные ошибки коэффициентов регрессии S_b, построить доверительные интервалы $b^- \le \beta \le b^+$ для истинных значений коэффициентов регрессии, соответствующие заданной надёжности гамма**.**



6. Проверить значимость коэффициентов регрессии по уровню альфа, построить вектор значимости и вектор значимых коэффициентов регрессии $b_{зн}$. Построить тренд и ошибки по значимым коэффициентам регрессии (если нет значимых – то использовать все незначимые). Как при этом изменится коэффициент детерминации $R_{зн}^2$?
7. Установить отсутствие гетероскедастичности по уровню значимости α и отсутствие автокорреляции по уровню значимости 0,05.
8. Построить графики $(\tilde{o}_1,\hat{y});(x_2,\hat{y});...(x_m,\hat{y})$ и подобрать хотя бы одну нелинейную инструментальную переменную $z=\varphi(x_j)$, повышающую коэффициент детерминации.
9. Построить линейную факторную модель по методу главных координат. Построить $\hat{Y},e,R_0^2$ по факторным переменным.



10. Выбрать главные факторы, объясняющие не менее 75% изменчивости наблюдаемых переменных. Построить $\hat{Y},e,R_0^2$ по главным факторным переменным. По факторным нагрузкам понять смысл первого главного фактора.

11. Выбрать 2 первых главных фактора и построить по ним $\hat{Y}, e, R_0^2$. На факторной плоскости (f_1, f_2) построить наблюдаемые объекты и разбить их на 2-3 класса по методу расстояний.

Входные данные

Измеренные факторы для статистической обработки в РГР

У факторы			Х факторы									
1	2	3	1	2	3	4	5	6	7	8	9	10
9,26	204,2	13,26	0,89	0,34	1,73	0,31	166,2	167,29	10,08	17,22	9889	0,28
9,44	209,6	10,16	0,93	0,33	0,99	0,15	186,1	92,88	14,76	18,39	2212	0,25
12,11	223,54	13,72	1,33	0,17	1,73	0,14	220,5	159,01	6,45	26,46	1078	0,47
10,81	236,7	12,83	0,68	0,32	0,47	0,18	169,3	93,96	21,83	22,37	1072	1,53
9,33	62	10,63	0,89	0,36	1,73	0,31	39,93	173,88	11,94	28,13	5526	0,21
9,87	53,1	9,12	1,53	0,33	1,33	0,17	40,41	162,3	12,6	17,55	4532	0,13
8,17	172,1	25,95	1,12	0,15	0,97	0,26	103	88,56	11,52	21,79	1265	0,38
9,12	56,5	23,39	0,99	0,32	1,82	0,29	37,02	101,16	8,28	19,52	5756	0,38
5,88	52,6	14,68	1,65	0,31	0,68	0,26	45,94	167,29	11,52	23,85	1182	0,2
6,3	46,6	10,05	0,56	0,15	1,8	0,28	40,07	140,76	32,4	21,88	6436	0,35
6,19	53,2	13,89	0,58	0,17	1,19	0,25	45,44	128,52	11,52	25,68	6964	0,2
5,46	30,1	9,68	1,53	0,15	0,97	0,49	41,08	177,84	17,28	18,13	4984	0,17
6,5	146,4	10,03	0,7	0,16	1,15	0,26	136,1	114,48	16,2	25,74	2249	0,25
6,61	18,1	9,13	1,77	0,15	0,02	0,28	42,39	93,24	13,36	21,21	6920	0,16
4,32	13,6	5,37	0,74	0,17	0,06	0,17	37,39	126,72	17,28	22,86	5736	0,21
7,37	89,8	9,86	1,08	0,34	1,39	0,17	101,8	91,27	9,72	16,38	4726	0,19
7,02	62,5	12,62	1,15	0,34	0,08	0,31	47,91	69,12	16,2	13,21	7208	1,24
8,25	46,3	5,02	0,97	0,34	0,77	0,18	32,61	66,24	24,88	14,41	8370	0,43
8,15	103,47	21,18	1,12	0,19	0,77	0,31	103,7	67,16	14,76	13,44	1076	0,14
8,72	73,3	25,17	0,99	0,19	1,08	0,18	38,95	50,4	7,56	13,69	6592	0,29
6,64	76,6	19,4	0,58	0,34	0,93	0,31	81,32	70,89	8,64	16,66	9981	0,43
8,1	73,01	21	1,03	0,34	0,1	0,15	67,75	72	8,64	15,06	7568	0,17
5,52	32,3	6,57	1,24	0,15	0,11	0,28	59,66	97,2	9	20,,09	4419	0,21
9,37	198,54	14,19	0,89	0,19	1,44	0,18	107,8	80,28	14,76	15,91	2089	0,42
13,17	598,12	15,81	0,68	0,34	0,48	0,14	512,6	51,48	10,08	18,27	2894	1,19
6,67	71,69	5,2	1,03	0,19	1,24	0,18	53,53	105,12	14,76	14,44	7468	1,87
5,68	90,63	7,96	0,73	0,32	0,77	0,29	80,83	128,52	10,38	22,88	8631	0,15
5,19	82,1	17,5	0,73	0,19	0,93	0,3	59,42	94,68	14,76	15,5	3131	0,03
10,02	76,2	17,16	0,85	0,33	0,13	0,27	36,96	85,32	20,52	19,35	6475	0,24
8,16	119,47	14,54	1,03	0,34	1,73	0,14	91,88	76,32	14,46	16,95	8206	0,93
3,78	21,83	6,21	0,47	0,36	0,77	0,29	17,16	153	24,88	30,53	4467	0,13
6,45	48,4	12,08	0,56	0,33	0,16	0,44	27,29	107,34	11,16	17,78	6518	0,27
10,38	173,5	9,39	0,89	0,32	0,74	0,14	184,3	90,72	6,45	22,09	2269	0,17
7,65	74,1	9,28	0,99	0,15	1,95	0,29	58,42	82,44	9,72	18,29	6810	0,24
8,77	68,6	11,44	1,95	0,16	0,58	0,18	59,31	79,12	3,24	26,05	6561	0,19
7	60,8	10,31	1,03	0,16	1,77	0,44	49,87	120,96	6,45	26,2	1273	0,29
11,06	355,6	8,65	0,01	0,2	0,7	0,31	391,3	84,6	5,4	17,26	7919	0,25
9,02	264,81	10,88	0,02	0,15	0,74	0,18	258,6	85,32	6,12	18,95	1431	0,36
13,28	526,62	9,87	0,6	0,33	1,15	0,14	75,14	101,52	8,64	19,66	9277	0,17
9,27	118,6	6,14	0,97	0,33	1,19	0,31	123,2	107,34	11,94	16,97	1220	0,23

Варианты заданий РГР

Nвар	Y	n	m	X1	X2	X3	X4	X5	X6	X7	X8	Альфа	Гамма
1	1	11	8	N	N+2	N+3	N+4	N+6	N+7	N+8	N+9	0,01	0,9
2	2	12	8	N	N+1	N+3	N+4	N+5	N+7	N+8	N+9	0,025	0,925
3	3	13	8	N	N+1	N+2	N+4	N+5	N+6	N+8	N+9	0,05	0,95
4	2	14	8	N	N+1	N+2	N+3	N+5	N+6	N+7	N+9	0,075	0,975
5	1	15	8	N	N+1	N+2	N+3	N+4	N+6	N+7	N+8	0,1	0,99
6	2	16	8	N	N+2	N+3	N+4	N+5	N+7	N+8	N+9	0,01	0,9
7	3	17	8	N	N+1	N+3	N+4	N+5	N+6	N+8	N+9	0,025	0,925
8	2	18	7	N	N+2	N+4	N+6	N+7	N+8	N+9		0,05	0,95
9	1	19	7	N	N+1	N+3	N+5	N+7	N+8	N+9		0,075	0,975
10	2	20	7	N	N+1	N+2	N+4	N+6	N+8	N+9		0,1	0,99
11	3	21	7	N	N+1	N+2	N+3	N+5	N+7	N+9		0,01	0,9
12	2	22	7	N	N+1	N+2	N+3	N+4	N+6	N+8		0,025	0,925
13	1	23	7	N	N+2	N+3	N+4	N+5	N+7	N+9		0,05	0,95
14	2	24	7	N	N+1	N+3	N+4	N+5	N+6	N+8		0,075	0,975
15	3	25	6	N	N+4	N+6	N+7	N+8	N+9			0,1	0,99
16	2	26	6	N	N+1	N+5	N+7	N+8	N+9			0,01	0,9
17	1	27	6	N	N+1	N+2	N+6	N+8	N+9			0,025	0,925
18	2	28	6	N	N+1	N+2	N+3	N+7	N+9			0,05	0,95
19	3	29	6	N	N+1	N+2	N+3	N+4	N+8			0,075	0,975
20	2	30	6	N	N+2	N+3	N+4	N+5	N+9			0,1	0,99
21	1	31	6	N	N+1	N+3	N+4	N+5	N+6			0,01	0,9
22	2	32	5	N	N+4	N+6	N+8	N+9				0,025	0,925
23	3	33	5	N	N+1	N+5	N+7	N+9				0,05	0,95
24	2	34	5	N	N+2	N+6	N+8	N+9				0,075	0,975
25	1	35	5	N	N+1	N+3	N+7	N+9				0,1	0,99
26	2	36	5	N	N+1	N+2	N+4	N+8				0,01	0,9
27	3	37	5	N	N+2	N+3	N+5	N+9				0,025	0,925
28	2	38	5	N	N+1	N+3	N+4	N+5				0,05	0,95
29	1	39	4	N	N+4	N+8	N+9					0,075	0,975
30	2	40	4	N	N+1	N+5	N+9					0,1	0,99
			4	N	N+2	N+6	N+9					0,01	0,9
			4	N	N+3	N+7	N+9					0,025	0,925
			4	N	N+1	N+4	N+8					0,05	0,95
			4	N	N+2	N+5	N+9					0,075	0,975
			4	N	N+1	N+3	N+5					0,1	0,99

Идентификационный номер группы N=ОСТАТ(Nгр/10), то есть последняя цифра в номере группы. **Nвар** номер варианта по **списку в журнале преподавателя**.

Приложения.

1. Полезные статистические функции Excel

$n = \times ÑÒÐÎ Ê(x_B)$. $n = \times ÑÒÎ ËÁ(x_B)$. $\tilde{o}_{\hat{A}} = \frac{1}{n}\sum_{i=1}^{n} x_i = ÑÐÇÍ À\times(x_B)$;

$D_{\hat{A}} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \overline{x}_B)^2 = ÄÈÑÏ .Ã(\tilde{o}_{\hat{A}})$; $S^2 = \frac{n}{n-1} D_B = ÄÈÑÏ .Â(\tilde{o}_{\hat{A}})$;

$\sigma_{\hat{A}} = \sqrt{D_B} = ÑÒÀÍ ÄÎ ÒÊËÎ Í .Ã(\tilde{o}_{\hat{A}})$; $S = \sqrt{S^2} = ÑÒÀÍ ÄÎ ÒÊËÎ Í .Â(\tilde{o}_{\hat{A}})$;

$f(x) = Í Î ÐÌ .ÐÀÑÏ (x, m, \sigma, 0)$, $F(x) = Í Î ÐÌ .ÐÀÑÏ (x, m, \sigma, 1)$,
$x_p = Í Î ÐÌ .Î ÁÐ(p, m, \sigma)$, $x_\alpha = Í Î ÐÌ .Î ÁÐ.Ï Õ(\alpha, m, \sigma)$.

$f(\chi_n^2) = ÕÈ 2.ÐÀÑÏ (\chi^2, n, 0)$, $F(\chi_n^2) = ÕÈ 2.ÐÀÑÏ (\chi^2, n, 1)$,
$\chi_p^2 = ÕÈ 2.Î ÁÐ(p, n)$, $\chi_\alpha^2 = ÕÈ 2.Î ÁÐ.Ï Õ(\alpha, n)$.

$f(t_n) = ÑÒÜÞ ÄÅÍ Ò.ÐÀÑÏ (t, n, 0)$, $F(t_n) = ÑÒÜÞ ÄÅÍ Ò.ÐÀÑÏ (t, n, 1)$,
$t_p = ÑÒÜÞ ÄÅÍ Ò.Î ÁÐ(p, n)$, $t_\alpha = ÑÒÜÞ ÄÅÍ Ò.Î ÁÐ.Ï Õ(\alpha, n)$
$t_{\alpha/2} = ÑÒÜÞ ÄÅÍ Ò.Î ÁÐ.2Õ(\alpha, n)$. $\varepsilon_\gamma = ÄÎ ÂÅÐÈ Ò(\gamma, s, n)$

$f(F) = F.ÐÀÑÏ (F, n1, n2, 0)$, $F_{ðàñï}(F) = F.ÐÀÑÏ (F, n1, n2, 1)$,
$F_p = F.Î ÁÐ(p, n1, n2)$, $F_\alpha = F.Î ÁÐ.Ï Õ(\alpha, n1, n2)$,

$r_{xy} = ÊÎ ÐÐÅË(x_B, y_B)$, $\hat{y} = ÒÅÍ ÄÅÍ ÖÈß(x_B, y_B)$, $S_{xy} = ÑÒÎ Ø(x_B, y_B)$,
$b_1 = Î ÒÐÅÇÎ Ê(x_B, y_B)$, $b_2 = Í ÀÊËÎ Í (x_B, y_B)$,

$\sum_{i=1}^{n} x_i = ÑÓÌ Ì (x_B, ..)$, $\sum_{i=1}^{n} x_i^2 = ÑÓÌ Ì ÊÂ(x_B)$, $\prod_{i=1}^{n} x_i = Ï ÐÎ ÈÇÂÅÄ(x_B, ..)$
$\sum_{i=1}^{n} x_i y_i = ÑÓÌ Ì Ï ÐÎ ÈÇ(x_B, y_B)$, $\sum_{i=1}^{n}(x_i - y_i)^2 = ÑÓÌ Ì ÊÂÐÀÇÍ (x_B, y_B)$,
$\sum_{i=1}^{n}(x_i^2 - y_i^2) = ÑÓÌ Ì ÐÀÇÍ ÊÂ(x_B, y_B)$, $\sum_{i=1}^{n}(x_i - \overline{x})^2 = ÊÂÀÄÐÎ ÒÊË(x_B)$.
$C = ÅÑËÈ("ËÎ ÃÈ\times"; À; Â)$ логический выбор значения.

$C = ÒÐÀÍ ÑÏ (À)$, $C = Ì Î ÁÐ(À)$, $C = Ì Î Ï ÐÅÄ(À)$, $C = Ì ÓÌ Í Î Æ(À; Â)$. При матричном результате засылка результата $C \Leftarrow Ctrl + Shift + Enter$.

2. Распределение Дарбина-Уотсона

$\alpha = 0.05$

n	$k^1 = 1$		$k^1 = 2$		$k^1 = 3$		$k^1 = 4$		$k^1 = 5$	
	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U	d_L	d_U
6	0,61	1,40	–	–	–	–				
7	0,70	1,36	0,47	1,90	–	–				
8	0,76	1,33	0,56	1,78	0,37	2,29				
9	0,82	1,32	0,63	1,70	0,46	2,13				
10	0,88	1,32	0,70	1,64	0,53	2,02				
11	0,93	1,32	0,66	1,60	0,60	1,93				
12	0,97	1,33	0,81	1,58	0,66	1,86				
13	1,01	1,34	0,86	1,56	0,72	1,82				
14	1.05	1.35	0.91	1.55	0.77	1.78				
16	1,10	1,37	0,98	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,13	1,38	1,02	1,54	0,90	1,71	0,78	1,90	0,67	2,10
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,40	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,20	1,41	1,10	1,54	1,00	1,68	0,90	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,96	1,80	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,90	1,92
24	1,27	1,45	1,19	1,55	1,10	1,66	1,01	1,78	0,93	1,90
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,30	1,46	1,22	1,55	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,10	1,75	1,03	1,85
29	1,34	1,48	1,27	1,56	1,20	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83

Библиография

1. Андерсен, Т. Введение в многомерный статистический анализ /Т. Андерсен. Москва, Наука, 1963, 498с.

2. Кендал, М. Статистические выводы и связи. Том 2 / М. Кендал, А. М. Стьюарт. Москва, Наука, 1973, 895с.

3. Гайдышев, И. Анализ и обработка данных: специальный справочник / И. Гайдышев. Санкт-Петербург, Питер, 2001, 752с.

4. Гмурман, В.Е. Теория вероятностей и математическая статистика / В.Е. Гмурман. Москва, Высшая школа, 2001, 408с.

5. Кремер, Н.Ш. Теория вероятностей и математическая статистика / Н.Ш. Кремер. Москва, ЮНИТИ, 2004, 573с.

6. Горбиков, С.П. Лекции по теории вероятностей и математической статистике: учебное пособие /С.П. Горбиков, Л.В. Филатов. Нижегор. гос. архитектур.-строит. ун-т., Нижний Новгород. ННГАСУ, 2011, 104с.

7. Вентцель, Е.С. Теория вероятностей и её инженерные приложения /Е.С. Вентцель, Л.А. Овчаров. Москва, Наука, 1988, 368с.

8. Смирнов, Н.В. Курс теории вероятностей и математической статистики для технических приложений / Н.В. Смирнов, И.В. Дунин-Барковский, Москва, Наука, 1969, 512с.

9. Корн, Г. Справочник по математике / Г. Корн, Т. Корн. Москва, Наука, 1977, 832с.

10. Методика статистической обработки эмпирических данных/ РТМ 44-62. Москва, Госстандарт, 1966, 101с.

11. Филатов Л.В. Проверка статистических гипотез / Л.В. Филатов. Нижегор. гос. архитектур.-строит. ун-т., Нижний Новгород. :ННГАСУ, 2003, 41с.

Содержание

Филатов Леонид Владимирович

ЗАДАЧИ СТАТИСТИЧЕСКОГО АНАЛИЗА В СТРОИТЕЛЬСТВЕ

Корреляционный, регрессионный и факторный анализ

Учебно-методическое пособие

Редактор
Сидоренко П.В.

Федеральное государственное бюджетное образовательное учреждение высшего образования «Нижегородский государственный архитектурно-строительный университет»
603950, Нижний Новгород, ул. Ильинская, 65.
Подписано в печать _____. Формат 60x90 1/16. Бумага газетная.
Печать трафаретная. Уч-изд. л.4,1. Усл.печ.л.4,2. Тираж 300экз. Заказ № _____
Полиграфический центр ННГАСУ, 603950, г.Н.Новгород, ул.Ильинская,65
http://www. nngasu.ru, srec@nngasu.ru