ФЕДЕРАЛЬНОЕ АГЕНТСТВО ПО ОБРАЗОВАНИЮ
Государственное образовательное учреждение высшего профессионального образования
"ТОМСКИЙ ПОЛИТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ"

Ю. Я. Кацман

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Методические указания к лабораторным работам
(Цикл лабораторных работ)

Издательство ТПУ
Томск 2008

УДК 519.22(07.07)

Составитель: Ю.Я. Кацман

Методические указания к лабораторным работам (цикл лабораторных работ) по дисциплине "Статистическая обработка экспериментальных данных". Томск: Изд. ТПУ, 2008.– 37с.

Методические указания разработаны для магистрантов, обучающих по программе: " Компьютерный анализ и интерпретация данных". Целью данной работы является изучение основных статистических методов анализа и интерпретации экспериментальных (случайных) данных с использованием математического пакета STATISTICA. В методических указаниях по каждой лабораторной работе приведены краткие теоретические сведения, варианты заданий и контрольные вопросы для самопроверки. В работе приведены примеры решения задач с помощью пакета STATISTICA.

Рекомендовано к печати Редакционно-издательским Советом
Томского политехнического университета

Рецензент: Останин С.А., кандидат технических наук, доцент кафедры программирования ФПМК ТГУ.

© Томский политехнический университет, 2008
© Оформление. Издательство ТПУ, 2008
© Ю.Я. Кацман, 2008

СОДЕРЖАНИЕ

Введение 3
Лабораторная работа № 1. 4
Первичная обработка эмпирических данных 4
Лабораторная работа № 2. 12
Проверка статистических гипотез 12
Лабораторная работа № 3. 24
Решение задачи линейного корреляционного и регрессионного анализа 24

Введение

Математизация знаний, опирающаяся на мощную техническую поддержку в виде современных ЭВМ, привела к широкому применению математико-статистических методов в работе специалистов. Дисциплина "Статистическая обработка экспериментальных данных" особенно необходима специалистам, деятельность которых связана с компьютерной обработкой данных: студентам, инженерам, магистрантам, аналитикам...

Для анализа данных и представления полученных результатов используется специальный пакет STATISTICA. Следует учесть, что использование этого пакета требует глубоких теоретических знаний статистических методов, умения строить статистические модели, корректировать параметры модели и анализировать полученные результаты.

Цикл лабораторных работ состоит из 5 индивидуальных заданий. Все работы выполняются на персональных компьютерах с ОС Windows. Статистический анализ данных проводится с использованием пакета STATISTICA 6.0 [1].

По каждой лабораторной работе необходимо представить отчет, который должен включать следующие пункты:

- Постановку задачи и цель исследований.
- В отчете необходимо привести исходные данные и результаты анализа, полученные при использовании различных модулей пакета.
- Данные анализа необходимо проиллюстрировать таблицами и графиками (используя пакет STATISTICA).

При успешном выполнении задания и правильном оформлении отчета студент (магистрант) допускается к защите лабораторной работы. Защита работы предусматривает знание всех изученных статистических методов по конкретной теме.

Лабораторная работа № 1.

Первичная обработка эмпирических данных

Цель работы – ознакомиться с простейшими приемами статистической обработки результатов наблюдений: группирование данных, получение выборочных характеристик, нахождение доверительных интервалов при заданном уровне значимости.

1. Теоретический обзор

Вспомним основные определения и понятия:

Выборкой объемом n для данной случайной величины ξ называется последовательность $x_1, x_2, \ldots, x_n$ независимых наблюдений этой величины.

Вариантами называют наблюдаемые значения x_i.

Вариационным рядом называется последовательность вариант, записанных в возрастающем порядке**.**

Частота обозначается ν_i и равна числу наблюдений варианты x_i.

Объём выборки равен $n = \sum_{i=1}^{k} \nu_i$, где $к -$ число различных значений вариант, наблюдаемых в опыте.

Относительными частотами (частостями) называется отношение соответствующих частот к объему выборки: $\omega_i = \nu_i / n$.

Статистической функцией распределения случайной величины X называется функция, определяющая для каждого значения x относительную частоту события $X < x$

$$P(X < x) = F^*(x) = \frac{\nu_x}{n} = \omega_x. \tag{1.1}$$

Показано (теорема Гливенко), что при $n \to \infty$ статистическая функция распределения стремится по вероятности к интегральной функции распределения: $F_n^*(x) \xrightarrow{p} F(x)$, а свойства $F_n^*(x)$ аналогичны свойствам $F(x)$.

Если выборка достаточно велика, то построенный на ее основе вариационный ряд неудобен для дальнейшего статистического анализа. В этом случае строится так называемый ***группированный статистический ряд***.

Малой выборкой называется такая выборка, при обработке которой методами, основанными на группировании наблюдений, нельзя достичь заданных точности и достоверности.

Большой считают такую выборку, при обработке которой можно перейти к группированию наблюдений без ощутимой потери информации и достижением заданных значений точности и достоверности.

При группировании данных соблюдаются определенные правила:

1. Объем выборки должен быть достаточно велик $(n \geq 50)$.
2. Число интервалов группирования m (число групп) должно находиться в интервале $5 \leq m \leq 20$.

3. Необходимо, по возможности, охватывать всю область данных.

4. Интервалы группирования не должны перекрываться. Не должно возникать никаких сомнений относительно того, в какой интервал попадает любое значение.

Существует множество различных формул для определения оптимального числа групп m выборки объемом n, приведем одну из них – формулу "Стерджесса":

$$m = 1 + \log_2 n. \tag{1.2}$$

Построив гистограмму относительных частот (частот) – аналог плотности распределения, мы сможем оценить вид распределения эмпирической выборки.

На следующем этапе анализа данных оценим числовые (точечные) характеристики выборки. Однако не забудем, что для установления качества или "правильности" любой оценки будем использовать *свойства* (требования) "*хороших оценок"*: **несмещенность**, **эффективность** и **состоятельность**.

Числовые характеристики эмпирического распределения называются **выборочными характеристиками**. Рассмотрим некоторые из них:

- *выборочное среднее*

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i; \tag{1.3}$$

- *выборочная дисперсия* (несмещённая) и *среднее квадратическое отклонение*

$$S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2; \tag{1.4}$$

$$S = \sqrt{S^2}; \tag{1.5}$$

- *выборочный коэффициент асимметрии*

$$Sk = \frac{\mu_3}{S^3}; \qquad \mu_3 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^3; \tag{1.6}$$

- *выборочный коэффициент эксцесса*

$$Ex = \frac{\mu_4}{S^4} - 3; \qquad \mu_4 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^4; \tag{1.7}$$

Вычисленные характеристики не позволяют судить о степени близости выборочных значений к оцениваемому параметру. Более предпочтительная процедура – построения интервала, который накрывает оцениваемый параметр с известной степенью достоверности. Такой подход называется "***интервальным оцениванием***".

Рассмотрим искомую процедуру. Пусть для параметра a получена *несмещённая* оценка $\tilde{a}$. Оценим возможную при этом ошибку. Назначим достаточно большую вероятность β (например: $\beta = 0.95, 0.97, 0{,}99$ ...), такую, что событие с вероятностью β можно считать практически достоверным. Теперь найдем такое значение ε, для которого выполняется соотношение

$$P\left(\left|\tilde{a} - a\right| < \varepsilon\right) = \beta. \quad (1.8)$$

Выразим диапазон возможных значений ошибки, обусловленный заменой a на $\tilde{a}$, в явном виде, причем, ошибки большие по абсолютной величине ε будут появляться с малой вероятностью $\alpha = 1 - \beta$:

$$P(\tilde{a} - \varepsilon < a < \tilde{a} + \varepsilon) = \beta. \quad (1.9)$$

Таким образом, с вероятностью β неизвестное значение параметра a попадает в интервал

$$I_\beta = (\tilde{a} - \varepsilon;\ \tilde{a} + \varepsilon\ ). \quad (1.10)$$

Вероятность β принято называть ***доверительной вероятностью,*** а интервал I_β – ***доверительным интервалом.***

Считая эмпирическую выборку объема n распределенной по нормальному закону, построим доверительные интервалы для математического ожидания и дисперсии:

a) доверительный интервал для математического ожидания нормального распределения $N(m_x, \sigma_x)$ при известной дисперсии определяется следующим образом.

Рассмотрим статистику $U = \frac{(\bar{x} - m_x)}{\sigma_x}\sqrt{n}$, имеющую нормальное распределение $N(0, 1)$. Следовательно, согласно (1.9) запишем

$$P\left(U_{\alpha/2} < U < U_{1-\alpha/2}\right) = \beta = 1 - \alpha, \quad (1.11)$$

где $U_{\alpha/2}$ и $U_{1-\alpha/2}$ – квантили стандартного нормального распределения $N(0, 1)$. Запишем неравенство (1.11), выполняющееся с вероятностью $1 - \alpha$ относительно m_x:

$$\bar{x} - \frac{\sigma}{\sqrt{n}}U_{\alpha/2} < m_x < \bar{x} + \frac{\sigma}{\sqrt{n}}U_{1-\alpha/2}. \quad (1.12)$$

Так как квантили нормального распределения связаны соотношением $U_{\alpha/2} = -U_{1-\alpha/2}$ и определяются по таблицам, окончательно получим

$$\varepsilon = U_{1-\alpha/2}\frac{\sigma}{\sqrt{n}};\ \ I_\beta(\bar{x} - \varepsilon, \bar{x} + \varepsilon). \quad (1.13)$$

b) доверительный интервал для математического ожидания нормального распределения $N(m_x, \sigma_x)$ при неизвестной дисперсии определяется следующим образом.

Так как дисперсия неизвестна, то непосредственно воспользоваться нормальным распределением $N(m_x, \sigma_x)$ нельзя. Однако известно, что случайная величина

$$t = \frac{(\bar{x} - m_x)}{S}\sqrt{n}, \quad (1.14)$$

где S –несмещенная оценка выборочного среднеквадратичного отклонения имеет распределение Стьюдента (t –распределение) с числом

степеней свободы $k = n - 1$. Для нахождения доверительного интервала потребуем, чтобы выполнялось равенство аналогичное (1.11):

$$P\left(\left|\frac{(\bar{x}-m_x)}{S}\sqrt{n}\right| < t\right) = \beta = 1 - \alpha, \qquad (1.15)$$

Величина t определяется по таблицам распределения Стьюдента для заданного уровня значимости (доверительной вероятности) и числа степеней свободы k. Квантили распределения Стьюдента связаны соотношением аналогичным нормальному распределению: $t_{\alpha/2} = -t_{1-\alpha/2}$. Запишем неравенство в выражении (1.15) относительно m_x:

$$\bar{x} - t_{\alpha/2,k}\frac{S}{\sqrt{n}} < m_x < \bar{x} + t_{1-\alpha/2,k}\frac{S}{\sqrt{n}}. \qquad (1.16)$$

Таким образом, для математического ожидания нормального распределения с неизвестной дисперсией, доверительный интервал определяется соотношением (1.15), а значение ε равно:

$$\varepsilon = t_{1-\alpha/2,k}\frac{S}{\sqrt{n}} \qquad (1.17)$$

c) Доверительный интервал для оценки дисперсии по выборочной дисперсии S^2 для нормального распределения строится аналогично выражению (1.11):

$$P(\sigma_1^2 < \sigma^2 < \sigma_2^2) = \beta = 1 - \alpha. \qquad (1.18)$$

Вспомним, что выборочная дисперсия и дисперсия нормального распределения связаны следующим соотношением:

$$kS^2 = \chi^2 \cdot \sigma^2, \qquad (1.19)$$

где случайная величина χ^2 –имеет "хи-квадрат" распределение с $k = n - 1$ степенями свободы. Отсюда следует, что квантили σ_1^2 и σ_2^2 будут определяться по таблицам распределения χ^2. Для заданной доверительной вероятности или, что тождественно, уровня значимости потребуем, чтобы площадь под кривой, лежащая левее левого квантиля, равнялась площади под кривой, расположенной правее правого квантиля, т.е.:

$$P(\chi^2 < \chi_1^2) = P(\chi^2 > \chi_2^2) = \frac{1-\beta}{2} = \frac{\alpha}{2}. \qquad (1.20)$$

Тогда из (1.19), учитывая (1.20), получим соответствующие границы интервала:

$$\sigma_1^2 = \frac{k \cdot S^2}{\chi_{1-\alpha/2,k}^2}; \quad \sigma_2^2 = \frac{k \cdot S^2}{\chi_{\alpha/2,k}^2}. \qquad (1.21)$$

2. Анализ данных в пакете Statistica 6.0

Первичную обработку эмпирических данных можно провести, используя данные (файл) из папки StatSoft\STATISTICA 6\Examples, однако мы создадим новый файл (выборку).

Создание файла данных

Запустим программу Statistica и последовательно выполним команды *File→New*. Во всплывшем меню *Create New Document* заполним поля *Number of variable* – 1; *Number of cases* – 125; As a stand-alone window. Будет создана пустая таблица (файл данных), состоящая из одного столбца и 125 строк. Документ можно сохранить – *Save as* Lab_1.sta. Заполним таблицу данными, распределенными по закону $N(m_x, \sigma_x)$. Для этого правой клавишей мыши щелкнем по имени переменной. Во всплывшем меню выбираем опцию *Variable Specs*…., затем в меню переменной в нижнем поле *Long name* … зададим вид функции *Functions* распределения случайных данных:

=VNormal(Rnd(1);5;3) $N(5,\ 3)$;

Можно задать другие законы распределения эмпирических данных, например:

=Rnd(100) равномерно распределенные на [0; 100];

= VExpon(Rnd(1);5) показательное распределение $\lambda = 5$.

Построение вариационного ряда

Для построения вариационного ряда нужно правой клавишей мыши щелкнуть по имени переменной и во всплывшем меню выбрать опцию *Sort Cases*. Не забудьте указать направление сортировки – от меньшего, к большему. При необходимости сохранить исходные данные, вариационный ряд можно построить в следующей переменной, предварительно скопировав в нее данные. К сожалению, анализировать вариационный ряд большой выборки достаточно сложно, поэтому применим группирование данных.

Группирование данных

В программе существуют различные модули для группирования данных и построения различных графиков. Прежде, чем группировать данные, качественно оценим, насколько наша выборка близка к нормальному распределению. С этой целью построим график на нормальной вероятностной бумаге. Выполним последовательно команды *Statistics→Basic Statistics/Tables→Descriptive Statistics→Normal probability plot*; Variable – Normal (см. рис. 1.1).

Для группирования данных воспользуемся командами *Graphs→Histograms*→2D *Histograms*. В открывшемся меню выберем опции Variables – Normal, Graph type – Regular, Fit type – Normal, Categories – 50 (число интервалов группирования). Опция Fit type строит на гистограмме частот теоретическую кривую, имеющую те же параметры, что и исходные данные. Построенные графики можно отредактировать и сохранить (см. рис. 1.2).

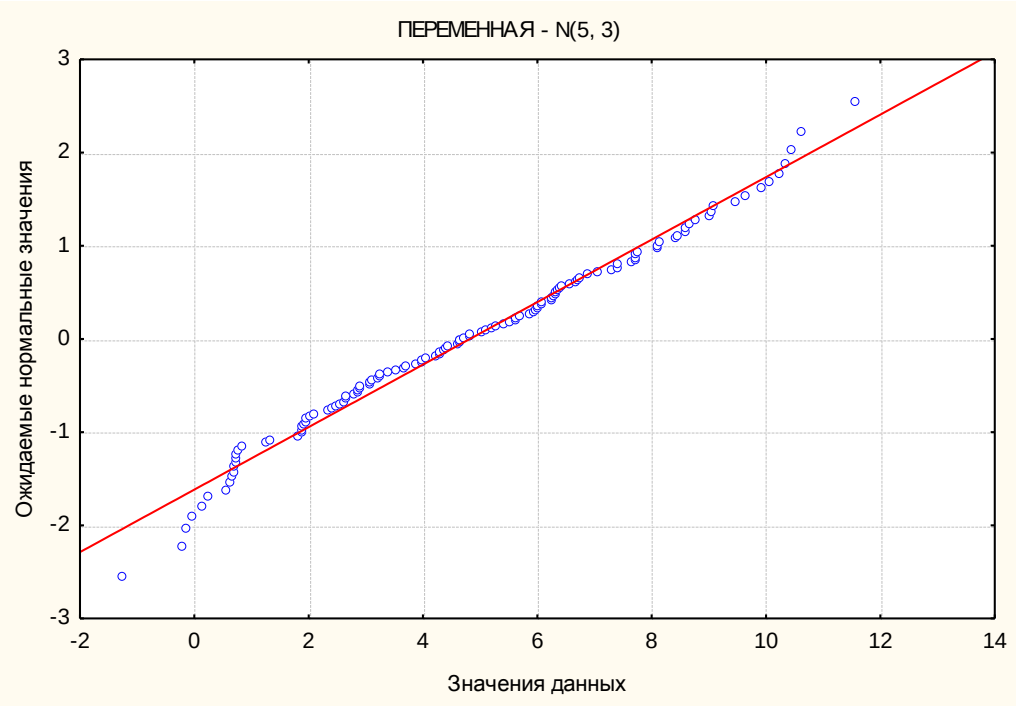


Рис. 1.1. График исходной выборки на нормальной вероятностной бумаге

При анализе графика следует учесть, чем ближе исходные данные к нормальному распределению, тем точнее они лягут на теоретическую прямую.

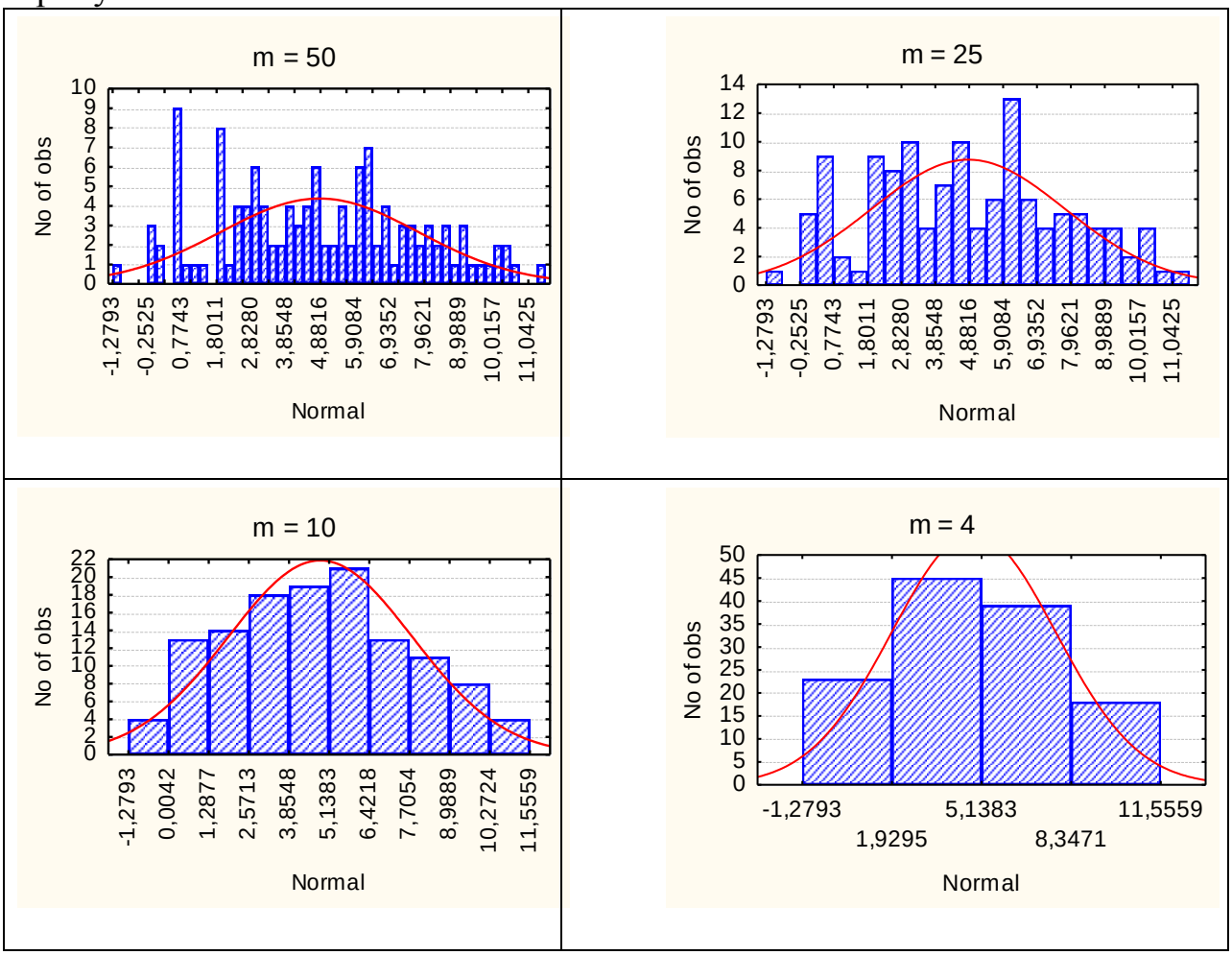


Рис. 1.2. Гистограмма частот (группированных)

На приведенных гистограммах (рис. 1.2) сплошной линией изображено нормальное распределение с параметрами равными выборочным характеристикам.

Числовые (точечные) характеристики выборки

Расчет характеристик выборки осуществим с помощью модуля *Basic Statistics/Tables* и процедуры этого модуля *Descriptive Statistics*. В открывшемся меню выберем имя переменной – Normal и перейдем на вкладку *Advanced*. Здесь можно выбрать интересующие нас характеристики, но, нажав клавишу *Select all stats*, выберем все. Отметим, что наряду с точечными характеристиками здесь рассчитываются границы доверительного интервала математического ожидания выборки при неизвестной дисперсии: *Interval* – 95%. По умолчанию доверительная вероятность равна 95 %, при необходимости этот параметр можно изменить. Все характеристики сведены в таблицу (рис. 1.3).

	Descriptive Statistics (Lab_1.sta)								
Variable	Valid N	Mean	Confidence -95,000%	Confidence +95,000%	Geometric Mean	Harmonic Mean	Median	Mode	Frequency of Mode
Normal	125	4,82088	4,30299	5,33877		7,97978	4,64561	Multipl	1

Рис. 1.3. Выборочные характеристики исходных данных

Интервальное оценивание

Так как процедуры нахождения доверительного интервала для математического ожидания при известной дисперсии и нахождения доверительного интервал для оценки дисперсии по выборочной дисперсии для данных, распределенных по нормальному закону, в пакете Statistica не реализованы, проведем эти расчеты вручную:

- *определение доверительный интервал для математического ожидания нормального распределения* $N(m_x, \sigma_x)$ *при известной дисперсии*;

Согласно выражению (1.13) нам необходимо определить квантиль распределения $N(0,1)$. Для этого воспользуемся калькулятором вероятности: *Statistica→Probability Calculator→Distributions*. В открывшемся окне выберем распределение Z (Normal), затем выберем опцию *Two-tailed*, а в окне *p*: – соответствующее значение доверительной вероятности и команду *Compute*. Соответствующее значение квантиля $U_{1-\alpha/2}$ получим в окне *X*:. При необходимости имеется возможность распечатать график распределения с соответствующими квантилями – *Create Graph*, *Send to Report*.

- *нахождение доверительного интервала для оценки дисперсии по выборочной дисперсии*;

Для нахождения доверительного интервала (1.21) необходимо найти квантили распределения $\chi^2_{1-\alpha/2,k}$ и $\chi^2_{\alpha/2,k}$. Как и ранее воспользуемся калькулятором вероятности и выберем распределение *Chi?* – “хи-квадрат”. В поле *df*: – число степеней свободы $k = n - 1$, в поле *p*: – соответствующее значение, равное половине уровня значимости $\alpha/2$ и команду *Compute*. Для

нахождения второго квантиля необходимо в поле *p*: – набрать значение равное $1-\alpha/2$ команду *Compute*. Второй квантиль можно найти, не изменяя поле *p*:, а выделив поля *Invers* и (1-*Cumulative p*), затем выполним команду *Compute*.

Теперь, используя инженерный калькулятор (Windows Калькулятор Плюс), по формулам (1.12) и (1.21) определим границы соответствующих интервалов.

3. Задание

1. Изучить основные модули системы Statistica 6.0.
   - Ознакомиться с графическими возможностями программы, визуализацией исходных данных и результатов анализа.
   - Научиться автоматически создавать отчет в системе Statistica.
2. Провести первичный статистический анализ случайных данных:
   - получить случайную выборку заданного объема с заданным законом распределения;
   - исследовать различные способы группирования данных;
   - вычислить (получить) основные выборочные (точечные) характеристики;
   - считая случайную выборку распределенной по нормальному закону, вычислить доверительные интервалы для математического ожидания и дисперсии при заданной доверительной вероятности.

Конкретные задания для каждого варианта приведены в табл. 1.1. В таблице принятыы следующие обозначений:

$N(m_x, \sigma_x)$ – гауссово распределение с соответствующим математическим ожиданием и средним квадратическим отклонением;

$R[l, u]$ – равномерное распределение на интервале от l до u;

$E(\lambda)$ – показательное (экспоненциальное распределение) с соответствующим параметром λ.

Таблица 1.1

№	Распределение	n	β	№	Распределение	n	β
1	N(5,3)	105	0.9	14	R[-5, -1]	160	0.83
2	R[1, 5]	110	0.91	15	E[0.333]	166	0.84
3	E[5]	125	0.92	16	N(-2,10)	175	0.85
4	N(2,10)	115	0.93	17	R[40, 100]	170	0.86
5	R[4, 10]	122	0.94	18	E[0.111]	177	0.87
6	E[0.2]	130	0.95	19	N(15,25)	134	0.88
7	N(15,2)	135	0.96	20	R[35, 60]	143	0.89
8	R[5, 20]	140	0.97	21	E[10]	177	0.9
9	E[1]	137	0.98	22	N(11,11)	144	0.91
10	N(12,1)	145	0.99	23	R[0, 1]	155	0.92
11	R[4, 15]	147	0.80	24	E[3.33]	180	0.93
12	E[0.1]	150	0.81	25	N(-5,1)	185	0.94
13	N(-5,3)	111	0.82	26	R[-5, 5]	190	0.95

4. Контрольные вопросы

1. Каковы основные задачи математической статистики?
2. Как связан объем выборки с возможностью группирования данных?
3. Как необходимо увеличить объем выборки для увеличения оптимального количества интервалов вдвое, согласно формуле "Стерджесса"?
4. Каковы свойства эмпирической функции распределения?
5. Какими свойствами обладают “хорошие оценки”?
6. Можно ли задать значение доверительной вероятности равным единице?
7. Как связан параметр λ с числовыми характеристиками показательного распределения?

Лабораторная работа № 2.

Проверка статистических гипотез

Цель работы – изучить основные методы проверки простых статистических (параметрических и непараметрических) гипотез, исследовать зависимость принятия (отклонения) нулевой гипотезы от значения уровня значимости.

1. Теоретический обзор

Приведем принцип ***практической уверенности***, лежащий в основе статистических методов проверки гипотез:

Если вероятность события A в данном испытании очень мала, то при однократном испытании можно быть уверенным в том, что событие A не произойдет, и в практической деятельности вести себя так, как будто, событие A вообще невозможно.

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

Статистическим критерием (или просто критерием) называют случайную величину K, которая служит для проверки гипотезы.

Нулевой (основной) называют выдвинутую гипотезу H_0. Вспомним основные определения и понятия:

Конкурирующей (альтернативной) называют гипотезу, которая противоречит выдвинутой нулевой гипотезе, и обозначается H_1.

При проверке гипотезы возможны два типа ошибок.

- Во-первых, гипотеза может быть отклонена, хотя фактически она верна. Такая ошибка называется ***ошибкой первого рода***.
- Во-вторых, гипотеза может быть принята, хотя фактически она неверна. Такая ошибка называется ***ошибкой второго рода***.

Мощность критерия равны значению $1-\beta$, где β – вероятность ошибки второго рода.

ПРОВЕРКА ПАРАМЕТРИЧЕСКИХ ГИПОТЕЗ

Гипотеза о равенстве (различии) двух выборочных средних

Пусть дана выборка из n_1 значений нормально распределённой СВ X и n_2 значений нормально распределенной СВ Y, причем

$$X \Rightarrow N(m_x, \sigma_x^2); \quad Y \Rightarrow N(m_y, \sigma_y^2).$$

Необходимо проверить гипотезу $H_0: \ m_x = m_y$, против гипотезы $H_1: \ m_x \neq m_y$ при заданном уровне значимости α.

Доказано, что выборочные средние $\bar{x}$ и $\bar{y}$ являются эффективными и несмещенными оценками соответствующих математических ожиданий с соответствующими дисперсиями: σ_x^2/n_1, σ_y^2/n_2. Однако на практике дисперсии σ_x^2 и σ_y^2 нам неизвестны, тогда воспользуемся объединённой оценкой $\sigma^2 \Rightarrow S^2$, полученной из обеих выборок:

$$S^2 = \frac{\sum_{i=1}^{n_1}(x_i-\bar{x})^2+\sum_{i=1}^{n_2}(y_i-\bar{y})^2}{n_1+n_2-2}. \qquad (2.1)$$

В качестве статистики (оценки) мы используем $t-$статистику (распределение Стьюдента) $k = n_1 + n_2 - 2$ степенями свободы:

$$t_k = \frac{\bar{x}-\bar{y}}{S(x-y)} = \frac{\bar{x}-\bar{y}}{\sqrt{\frac{S^2}{n_1}+\frac{S^2}{n_2}}} = \frac{\bar{x}-\bar{y}}{S\cdot\sqrt{\frac{n_1+n_2}{n_1\cdot n_2}}}. \qquad (2.2)$$

Теперь сравним полученное значение с соответствующим значением квантиля и, если $t_{k,1-\alpha/2} > t_k$, то принимаем нулевую гипотезу, в противном случае нулевая гипотеза отвергается, и считаем справедливой альтернативную гипотезу H_1.

Замечание *Для статистики Стьюдента можно рассматривать как двусторонние, так и односторонние критерии.*

Критерий Фишера

Критерий Фишера применяется при проверке гипотезы о равенстве дисперсий двух генеральных совокупностей, распределенных по нормальному закону. Гипотезы о дисперсиях возникают довольно часто, так как дисперсия характеризует такие исключительно важные показатели, как точность машин, приборов, технологических процессов, риск, связанный с отклонением доходности активов от ожидаемого уровня, и т.д.

Сформулируем задачу. Пусть имеются две нормально распределенные совокупности, дисперсии которых равны σ_1^2 и σ_2^2. Необходимо проверить нулевую гипотезу о равенстве дисперсий, т.е. H_0: $\sigma_1^2 = \sigma_2^2$ относительно конкурирующей H_1: $\sigma_1^2 > \sigma_2^2$ или H_1': $\sigma_1^2 \neq \sigma_2^2$.

Для проверки гипотезы H_0 из этих совокупностей взяты две независимые выборки объемом n_1 и n_2. Так как оценки дисперсий σ_1^2 и σ_2^2

нам неизвестны, воспользуемся несмещенными выборочными оценками дисперсий S_1^2 и S_2^2.

Тогда при справедливости гипотезы H_0: $\sigma_1^2 = \sigma_2^2 = \sigma^2$ в качестве оценки σ^2 можно взять те же дисперсии S_1^2 и S_2^2, рассчитанные по элементам первой и второй выборок.

Известно, что выборочные характеристики $\frac{(n_1-1)\cdot S_1^2}{\sigma^2}$ и $\frac{(n_2-1)\cdot S_2^2}{\sigma^2}$ имеют распределение χ^2 соответственно с $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$ степенями свободы, а их отношение $\frac{\frac{1}{k_1}\chi^2(k_1)}{\frac{1}{k_2}\chi^2(k_2)}$ имеет $F-$ распределение Фишера – Снедекора с k_1 и k_2 степенями свободы. Следовательно, случайная величина F, определяемая отношением:

$$F = \frac{\frac{1}{n_1-1}\left[(n_1-1)\cdot S_1^2/\sigma^2\right]}{\frac{1}{n_2-1}\left[(n_2-1)\cdot S_2^2/\sigma^2\right]} = \frac{S_1^2}{S_2^2}, \qquad (2.3)$$

т.е. отношение несмещенных выборочных дисперсий имеет $F-$ *распределение Фишера – Снедекора с* $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$ степенями свободы.

Очевидно, что при равенстве дисперсий величина критерия будет равна единице. В остальных случаях она будет больше (меньше) единицы. При формировании критерия отклонения (принятия) гипотезы H_0 следует учесть, что распределение статистики F (в отличие от нормального или распределения Стьюдента является несимметричным.)

Критерий Фишера $F(\alpha, k_1, k_2)$ – двусторонний критерий, и нулевая гипотеза $H_0: S_1^2 = S_2^2$ принимается (отвергается альтернативная гипотеза $H_1: S_1^2 \neq S_2^2$) если $F(\alpha/2, k_1, k_2) < F < F(1-\alpha/2, k_1, k_2)$.

ПРОВЕРКА НЕПАРАМЕТРИЧЕСКИХ ГИПОТЕЗ

Критерий согласия хи-квадрат

Одной из важнейших задач математической статистики является ***установление теоретического закона распределения случайной величины*** по опытному (эмпирическому распределению), представляющему вариационный ряд.

Предположения о ***виде закона распределения*** выдвигаются исходя из теоретических предпосылок, опыта аналогичных исследований, и, наконец, на основании графического изображения эмпирического распределения.

Параметры эмпирического распределения, как правило, неизвестны, поэтому их заменяют выборочными оценками (несмещенными, эффективными и состоятельными).

При сравнении теоретического и эмпирического распределения неизбежны расхождения. Естественно возникает вопрос, обусловлены ли эти

расхождения только случайными факторами, связанными с ограниченным числом наблюдений (объемом выборки), или они являются существенными и обусловлены неудачным выбором теоретического закона распределения. Для ответа на этот вопрос и служат ***критерии согласия***.

Критерий согласия хи-квадрат Пирсона является весьма общим методом построения тестов для проверки различных гипотез. Рассмотрим исходную схему.

Введем обозначения:

$A_1,\ A_2, \dots, A_m -\ m$ возможных исходов некоторого опыта;

$p_1, p_2, \dots, p_m -$ вероятности соответствующих исходов, причем $\sum_{i=1}^{m} p_i = 1$;

$n -$ число независимых повторений опыта;

$\nu_1, \nu_2, \dots, \nu_m -$ число появлений соответствующих исходов в n опытах, причем $\sum_{i=1}^{m} \nu_i = n$;

$p_1^0, \dots, p_m^0$ –гипотетические значения вероятностей, $p_i^0 > 0,\ \sum_{i=1}^{m} p_i^0 = 1$;

Требуется по наблюдениям $\nu_1, \nu_2, \dots, \nu_m$ проверить гипотезу H о том, что вероятности $p_1, p_2, \dots, p_m$ имеют значения $p_1^0, \dots, p_m^0$, т.е.
$H_0: p_i = p_i^0, i = 1, \dots m.$

Оценками для $p_1, p_2, \dots, p_m$ являются $\hat{p}_1 = \nu_1/n, \dots\ \hat{p}_m = \nu_m/n$. Мерой расхождения между гипотетическими и эмпирическими вероятностями принимается величина

$$X^2 = n\sum_{i=1}^{m} p_i^0 \left(\frac{\hat{p}_i - p_i^0}{p_i^0}\right)^2 = \sum_{i=1}^{m} \frac{\left(\nu_i - np_i^0\right)^2}{np_i^0} = \sum_{i=1}^{m} \frac{\nu_i^2}{np_i^0} - n. \quad (2.4)$$

Статистика X^2 называется статистикой хи-квадрат Пирсона, а процедура проверки гипотезы состоит в том, что если величина X^2 приняла “слишком большое значение”, т.е. $X^2 \geq h$, то гипотеза H_0 отклоняется. В противном случае будем говорить, что наблюдения не противоречат гипотезе. На вопрос, что означает “слишком большое значение”, отвечает следующая теорема.

Теорема К. Пирсона. Если гипотеза H верна и $p_i^0 > 0,\ i = 1, \dots m$, то при $n \to \infty$ распределение статистики X^2 асимптотически подчиняется распределению хи-квадрат с $m - 1$ степенями свободы, т.е.

$$P\{X^2 < x/H\} \ \to\ F_{m-1}(x) \equiv P\{\chi_{m-1}^2 < x\}. \quad (2.5)$$

Порог h выберем из условия: вероятность ошибки первого рода должна быть малой – равной выбираемому значению уровня значимости α:

$$P\{\text{отклонить}\, H/H \text{ верна}\} = P\{X^2 \geq h/H\} \cong P\{\chi_{m-1}^2 \geq h/H\} = \alpha,$$

откуда

$$h = \chi_{1-\alpha, m-1}^2 \quad (2.6)$$

– квантиль уровня $1 - \alpha$ распределения хи-квадрат с $m - 1$ степенями свободы.

На практике случайную выборку из n независимых наблюдений сгруппируем по $k-$ интервалам, называемым *интервалами группировки*. Число степеней свободы распределения хи-квадрат в этом случае равно k (число интервалов группировки) минус число различных независимых линейных ограничений, наложенных на наблюдения:

- Первое ограничение связано с тем, что частота в последнем интервале группировки полностью определяется частотами всех остальных интервалов, т.е. не является независимой величиной.
- Если гипотетическая (предполагаемая) плотность – нормальная, с неизвестным математическим ожиданием и дисперсией, то появятся два дополнительных ограничения, поскольку для подбора нормальной плотности мы используем две выборочные оценки $(\bar{x}, S^2)$.

Естественно, при проверке нормальности распределения $m = k - 3$.

Критерий Колмогорова

При анализе выборок малого объема невозможно применить критерий χ^2 (группирование данных некорректно). В этом случае часто используется критерий Колмогорова, в котором в качестве меры расхождения между теоретическим и эмпирическим распределениями рассматривают максимальное значение абсолютной величины разности между эмпирической функцией распределения $F_n(x)$ и соответствующей теоретической функцией распределения:

$$D = \max_{-\infty \le x \le \infty} \left| F_n(x) - F(x) \right| \tag{2.7}$$

Оценка D называется ***статистикой критерия Колмогорова***.

Доказано, что какова бы ни была функция распределения $F(x)$ непрерывной случайной величины X, при неограниченном увеличении числа наблюдений $n \to \infty$ вероятность неравенства $P(D\sqrt{n} \ge \lambda)$ стремится к пределу:

$$P(\lambda) = 1 - \sum_{k=-\infty}^{\infty} (-1)^k \cdot e^{-2k^2\lambda^2} \tag{2.8}$$

Задавая уровень значимости α, из соотношения $P(\lambda) = \alpha$ находим соответствующее критическое значение λ_α.

Рассмотрим схему применения критерия Колмогорова:

- На первом этапе строятся эмпирическая функция распределения $F_n(x)$ и предполагаемая теоретическая функция распределения $F(x)$
- На втором этапе определяется мера расхождения D между теоретическим и эмпирическим распределением по формуле (2.7) и вычисляется величина

$$\lambda = D\sqrt{n} \tag{2.9}$$

- Если вычисленное значение λ окажется больше критического λ_α, определенного на уровне значимости α, то нулевая гипотеза H_0 о том, что

случайная величина X имеет заданный закон распределения, отвергается (односторонний критерий). Если $\lambda < \lambda_\alpha$, то считаем, что гипотеза H_0 не противоречит опытным данным.

Двухвыборочный критерий Колмогорова – Смирнова

Рассматриваются две выборки случайных величин

$$X: x_1, x_2, \dots, x_{n_1} \text{ и } Y: y_1, y_2, \dots, y_{n_2}.$$

Перед исследователем стоит вопрос: обе выборки извлечены из совокупности с одним и тем же законом распределения вероятностей? Говоря языком математической статистики, необходимо проверить нулевую гипотезу $H_0: F_{n_1}(x) = F_{n_2}(y)$ о совпадении функций распределения вероятностей в двух выборках. В качестве меры расхождения здесь рассматривается разность двух эмпирических функций распределения вероятностей [2]:

$$D_n^* = \max\left|F_{n_1}(x) = F_{n_2}(y)\right|. \tag{2.10}$$

Статистика критерия Колмогорова-Смирнова имеет вид:

$$\lambda' = \sqrt{\frac{n_1 \cdot n_2}{n_1 + n_2}} \max\left|F_{n_1}(x) = F_{n_2}(y)\right|. \tag{2.11}$$

Гипотеза H_0 отвергается, если фактически наблюдаемое значение статистики λ' больше критического λ'_α, т.е. $\lambda' > \lambda'_\alpha$, и принимается в противном случае.

2. Проверка статистических гипотез в пакете Statistica 6.0

Проверка параметрических гипотез

Для проверки гипотез предварительно подготовим файл данных, включающий 5 переменных, распределенных по нормальному закону. Для каждой переменной проведено 25 наблюдений. Анализ данных проведем с помощью модуля *Basic Statistics/Tables* и процедуры этого модуля *Descriptive Statistics*. Для визуального (качественного) анализа построим графики типа “коробок” (“ящик с усами”), для чего выберем соответствующие переменные и нажмем клавишу *Box & whisker plot for all variables* см. рис.2.1.

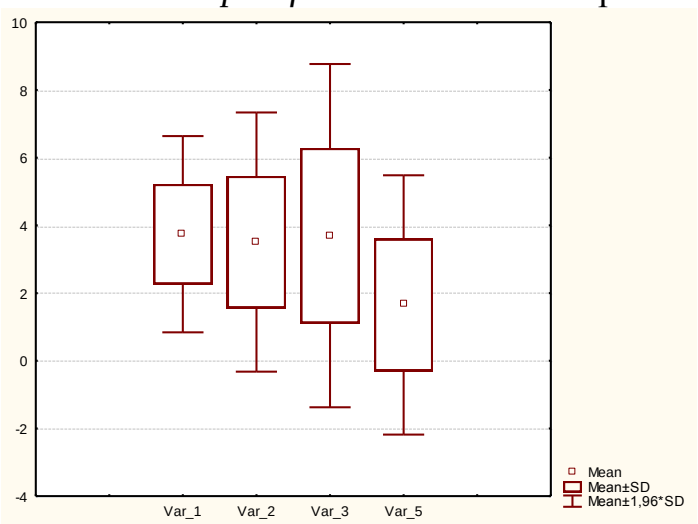


Рис. 2.1. Значения выборочных средних и стандартных отклонений

Анализ графиков свидетельствует о близости значений средних для переменных Var 1, Var 2 и Var 3. Что касается дисперсий (стандартных отклонений), то они близки для переменных Var 1, Var 2 и Var 5.

Для количественных оценок воспользуемся процедурой *t-test for independent samples* (критерий Стьюдента для независимых выборок) из модуля *Basic Statistics/Tables.* Затем выберем переменные Variables (groups), средние значения которых будем сравнивать. Теперь, чтобы запустить тест необходимо нажать клавишу *Summary* или *Summary: T-test*. По умолчанию мы работали на вкладке *Quick*, однако, перейдя на вкладку *Options*, мы можем в окошке *p-level for highlighting*: задать произвольный уровень значимости. Результаты теста представлены в виде таблицы. Мы последовательно сравнили три пары переменных, сведя полученные результаты в одну таблицу (рис. 2.2).

Group 1 vs. Group 2	T-test for Independent Samples (Lab_2.sta) Note: Variables were treated as independent samples										
	Mean Group 1	Mean Group 2	t-value	df	p	Valid N Group 1	Valid N Group 2	Std.Dev. Group 1	Std.Dev. Group 2	F-ratio Variances	P Variances
Var 1 vs Var 2	,7407	,5064	,4276	8	,6714	0	0	,47937	,95327	,74335	,23483
Var 1 vs Var 3	,7407	,6956	,0676	8	,9465	0	0	,47937	,58758	,05937	,01894
Var 1 vs Var 5	,7407	,6490	,8172	8	,0005	0	0	,47937	,95353	,74375	,23461

Рис. 2.2. Результаты сравнения выборочных средних и выборочных дисперсий

Прежде, чем анализировать полученные результаты, рассмотрим принятые в таблице обозначения:

Var 1 vs Var 2 – имена двух сравниваемых переменных (выборок);

Mean Group 1, Mean Group 2 – среднее арифметическое первой и второй выборки, соответственно;

t-value – значение статистики (распределение Стьюдента);

df – число степеней свободы распределения Стьюдента;

p – вероятность того, что случайная величина примет значения большее, чем t-value (двусторонний критерий);

Valid N Group 1, Valid N Group 2 – объем первой и второй выборки соответственно;

Std.Dev. Group 1, Std.Dev. Group 2 – стандартное отклонение первой и второй выборки соответственно;

F-ratio Variances – значение отношений дисперсий двух выборок;

P Variances – вероятность того, что случайная величина примет значение большее F.

Замечание *В программе Statistica при обращении к процедуре сравнения двух выборочных средних* (*критерий Стьюдента*) *одновременно реализована проверка равенства двух выборочных дисперсий* (*критерий Фишера*). *В программе* (*см. табл.2.2*) *критерий Фишера реализован таким образом, что максимальная дисперсия заносится в числитель, а меньшая в знаменатель статистики F.*

Анализ результатов

Процедуру проверки параметрической гипотезы и анализа полученных

результатов можно формализовать:

1. Исходные выборки должны быть распределены по нормальному закону – это условие выполнено.
2. При проверке критерия Стьюдента дисперсии (выборочные дисперсии) двух выборок должны быть равны. Для сравнения дисперсий используем критерий Фишера (2.3). Анализ результатов табл.2.2 (последние 2 столбца) свидетельствует, что при уровне значимости $\alpha = 0.05$ дисперсии первой и второй выборок, а также дисперсии первой и пятой выборок отличаются незначимо и принимается гипотеза H_0 – дисперсии равны. Таким образом, для этих пар выборок (строки 1 и 3 табл. 2.2) проверка t −критерия – корректна. При сравнении дисперсий первой и третей выборок (вторая строка таблицы) гипотезу H_0 можно принять с вероятностью менее 2%, что меньше уровня значимости. Поэтому, принимаем альтернативную гипотезу $H_1 -$ дисперсии различны. Для выборок с различными дисперсиями применение критерия Стьюдента не корректно, т.е. мы ничего не можем сказать, сравнивая выборочные средние переменных Var 1 и Var 3.
3. Проверка t −критерия для первой и второй переменных (первая строка таблицы) свидетельствует, что с вероятностью более 67% верна нулевая гипотеза – выборочные средние равны. Совершенно иная картина при сравнении средних первой и пятой выборок (третья строка таблицы). Можно считать их равными лишь с вероятностью ~ 0.05%, естественно при этом принять конкурирующую гипотезу – средние двух выборок не равны.

Проверка непараметрических гипотез

Для проверки критерия согласия Пирсона подготовим файл данных, включающий одну переменную с 500 наблюдений. Заполним переменную случайными данными, равномерно распределенными на интервале [10, 20]. Анализ данных проведем, используя модуль *Statistics* → *Distribution Fitting*. В открывшемся окне выберем (ожидаемый) вид распределения – Rectangular, OK. В новом открывшемся окне заполним минимум данных, оставив остальные по умолчанию: *Variable*: Var1. В окне *Distribution*: активизируется вид выбранного теоретического распределения. Работая на вкладке Quick, активизируем клавишу *Plot of observed and expected distribution*.

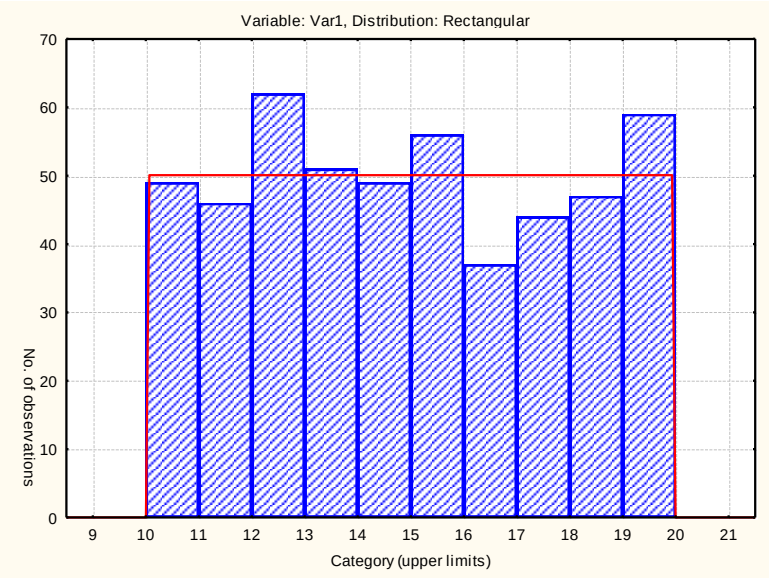


Рис. 2.3. Гистограмма частот случайной выборки

Активируем вторую клавишу *Summary*: *Observed and expected distribution.* Все результаты сведены в таблицу и приведены на рис. 2.4.

	Variable: Var1, Distribution: Rectangular (Lab_2_b.sta) Chi-Square = 10,05597, df = 7 (adjusted) , p = 0,18542								
Upper Boundary	Observed Frequency	Cumulative Observed	Percent Observed	Cumul. % Observed	Expected Frequency	Cumulative Expected	Percent Expected	Cumul. % Expected	Observed-Expected
<= 10,00000	0	0	0,0000	0,000	0,0000	0,000	0,0000	0,000	0,000
11,00000	49	49	9,8000	9,800	49,7737	49,773	9,9547	9,954	-0,773
12,00000	46	95	9,2000	19,000	50,0809	99,854	10,0161	19,970	-4,081
13,00000	62	157	12,4000	31,400	50,0809	149,935	10,0161	29,987	11,919
14,00000	51	208	10,2000	41,600	50,0809	200,016	10,0161	40,003	0,919
15,00000	49	257	9,8000	51,400	50,0809	250,097	10,0161	50,019	-1,081
16,00000	56	313	11,2000	62,600	50,0809	300,178	10,0161	60,035	5,919
17,00000	37	350	7,4000	70,000	50,0809	350,259	10,0161	70,051	-13,081
18,00000	44	394	8,8000	78,800	50,0809	400,340	10,0161	80,068	-6,081
19,00000	47	441	9,4000	88,200	50,0809	450,421	10,0161	90,084	-3,081
20,00000	59	500	11,8000	100,000	49,5786	500,000	9,9157	100,000	9,421
< Infinity	0	500	0,0000	100,000	0,0000	500,000	0,0000	100,000	0,000

```
Number of valid cases:500
Observed mean = 14,970339, Observed variance = 8,489621
Distribution: Rectangular
```

Рис. 2.4. Проверка критерия согласия Пирсона

В таблице приведены частоты и накопленные частоты для исходного и теоретического распределений. В заголовке таблицы указывается вид теоретического распределения, с которым сравниваются случайные данные, приводится статистика Пирсона и вероятность принятия нулевой гипотезы. Для нашей выборки вероятность принятия гипотезы H_0 более 18%, т.е. нет оснований отвергать гипотезу о том, что исходная выборка распределена равномерно. Обратим внимание, что в приведенной таблице, по умолчанию, выборка разбита на 12 групп. Перейдем на вкладку *Parameters* и установим: *Number of categories*: 10, *Lower limit*: 10, *Upper limit*: 20. Теперь таблица частот будет выглядеть корректнее. После этой коррекции вернемся на вкладку Quick и сравним исходные данные с нормальным распределением: *Distribution*: Normal.

Upper Boundary	Variable: Var1, Distribution: Normal (Lab_2_b.sta) Chi-Square = 39,47135, df = 7, p = 0,00000								
	Observed Frequency	Cumulative Observed	Percent Observed	Cumul. % Observed	Expected Frequency	Cumulative Expected	Percent Expected	Cumul. % Expected	Observed-Expected
<= 11,00000	49	49	9,8000	9,800	43,2484	43,248	8,6496	8,6497	5,7516
12,00000	46	95	9,2000	19,000	33,7500	76,998	6,7500	15,3997	12,2499
13,00000	62	157	12,4000	31,400	47,7246	124,723	9,5449	24,9446	14,2754
14,00000	51	208	10,2000	41,600	60,0554	184,778	12,0110	36,9557	-9,0554
15,00000	49	257	9,8000	51,400	67,2520	252,030	13,4504	50,4061	-18,2521
16,00000	56	313	11,2000	62,600	67,0197	319,050	13,4039	63,8101	-11,0198
17,00000	37	350	7,4000	70,000	59,4353	378,485	11,8870	75,6971	-22,4353
18,00000	44	394	8,8000	78,800	46,9061	425,391	9,3812	85,0784	-2,9061
19,00000	47	441	9,4000	88,200	32,9425	458,334	6,5885	91,6669	14,0575
< Infinity	59	500	11,8000	100,000	41,6656	500,000	8,3331	100,000	17,3343

```
Number of valid cases:500
Observed mean = 14,970339, Observed variance = 8,489621
Distribution: Normal
Parameters: Mean = 14,97034, Variance = 8,489621
```

Рис. 2.5. Критерий Пирсона для исходного и гауссового распределений

В приведенной таблице нет пустых групп, при сравнении исходного и нормального распределений значение статистики Пирсона очень велико, т.е. нулевую гипотезу можно принять с вероятностью менее $1 \cdot 10^{-3}$. Естественно отвергнуть нулевую гипотезу и принять альтернативную – **случайные данные распределены не по нормальному закону**.

Проверим гипотезу об однородности двух выборок – **критерий Колмогорова-Смирнова**.

Выберем модуль *Nonparametrics* из меню *Statistics*. Затем выберем пункт *Comparing two independent samples* (*groups*), Появится меню, предлагающее проверку пяти различных тестов. Нас интересует *Kolmogorov-Smirnov two-sample test.*

Замечание *При сравнении распределения двух выборок случайные данные двух выборок заносятся в одну переменную, а коды выборок (текстовые или числовые) заносятся во вторую переменную. При этом порядок заполнения данных безразличен, объемы сравниваемых выборок, как правило, различны*.

Подготовим исходные данные, задав в переменной *Uniform* 29 случайных значений, равномерно распределенных на интервале [10, 20]. Соответствующие коды указаны в переменной *Codes*. Первая выборка содержит 13, а вторая – 16 наблюдений. Заполнив поля переменных *Dependent variable list*: Uniform, *Indep*. (*grouping*) *variable*: Codes. Результаты проверки теста приведены на рис. 2.6.

variable	Kolmogorov-Smirnov Test (Lab_2_b.sta) By variable Codes Marked tests are significant at p <,05000								
	Max Neg Differnc	Max Pos Differnc	p-level	Mean Group 1	Mean Group 2	Std.Dev. Group 1	Std.Dev. Group 2	Valid N Group 1	Valid N Group 2
Uniform	-0,35096	0,09134	p > .10	14,0991	15,0416	2,41003	2,13435	13	16

Рис. 2.6. Критерий Колмогорова-Смирнова для двух выборок из одной генеральной совокупности

В таблице результатов приведены основные параметры двух выборок, максимальная положительная и отрицательная разность двух эмпирических

функций распределения и вероятность того, что верна гипотеза H_0. Т.к. эта вероятность достаточно велика (более 10%) при уровне значимости 5%, то нет оснований отвергнуть нулевую гипотезу.

Проверим еще раз критерий Колмогорова-Смирнова, задав в качестве второй выборки нормально распределенную случайную величину $N(10, 5)$. Данные двух выборок внесем в переменную Norm, соответствующие коды выборок оставим в переменной Codes. Результаты теста приведены на рис. 2.7.

Kolmogorov-Smirnov Test (Lab_2_b.sta)
By variable Codes
Marked tests are significant at p <,05000

variable	Max Neg Differnc	Max Pos Differnc	p-level	Mean Group 1	Mean Group 2	Std.Dev. Group 1	Std.Dev. Group 2	Valid N Group 1	Valid N Group 2
Norm	-0,06250	0,56250	p < .025	14,0991	9,93990	2,41003	6,33143	13	16

Рис. 2.7. Критерий Колмогорова-Смирнова для двух различно распределенных выборок

Теперь две эмпирические функции распределения можно считать одинаковыми с вероятностью менее 2.5%, что значительно меньше уровня значимости. В этом случае нулевая гипотеза отвергается, и принимаем альтернативную гипотезу – две выборки имеют различные функции распределения. В непохожести двух функций распределений можем убедиться визуально, для чего построим их. Предварительно 13 наблюдений равномерно распределенной случайной величины скопируем в переменную Rand, а 16 значений нормального распределения скопируем в переменную Gaus. Затем выполнимся модулем *Graphs* → *2D Histograms*. Далее выбираем опции: *Graph type* → Multiple; *Fit type* → Off; *Showing Type* → Standard; *Variables* →Rand-Gaus; *Categories* → 5; *Show* percentages; *Y axis:* N&%. Нажмем OK и получим результат (рис. 2.8).

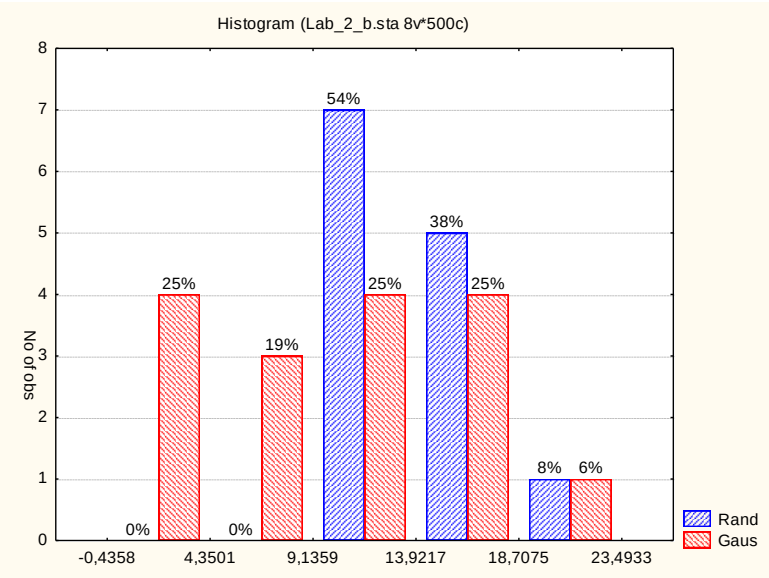


Рис. 2.8. Сравнение двух эмпирических функций распределения

3. Задание

1. Подготовка файла данных в соответствии с указанным вариантом (см. табл. 2.1).
2. Считая данные, распределенными по нормальному закону, выделите 2 выборки объемом k_1 и k_2.
3. Проверьте параметрические гипотезы (критерий Стьюдента, критерий Фишера), изменяя исходные выборки так, чтобы выборочные средние и выборочные дисперсии значимо/незначимо отличались.
4. Проверьте критерий согласия Пирсона для исходных данных.
5. Проверьте тест Колмогорова-Смирнова для двух выборок объемом k_1 и k_2; изменяя исходные данные обеспечьте равенство, неравенство эмпирических функций распределения.

Таблица 2.1

№	Распределение	n	k_1	k_2	№	Распределение	n	k_1	k_2
1	N(0,1)	225	10	13	14	R[-3, 0]	260	26	16
2	R[2, 6]	140	7	8	15	E[0.333]	366	36	17
3	E[7]	125	20	23	16	N(-2,1)	175	15	10
4	N(10,2)	315	15	15	17	R[50, 75]	270	17	27
5	R[3, 8]	222	10	15	18	E[0.111]	177	12	12
6	E[1/3]	130	12	17	19	N(15, 5)	234	11	14
7	N(1,2)	335	21	22	20	R[35, 70]	440	14	41
8	R[1, 10]	140	12	12	21	E[50]	277	27	22
9	E[0.1]	137	6	8	22	N(13,11)	244	16	18
10	N(2,1)	145	15	18	23	R[0, 10]	255	14	14
11	R[4, 5]	170	17	10	24	E[2.5]	180	18	17
12	E[0.4]	150	7	9	25	N(-10,10)	110	11	21
13	N(-5,5)	333	30	33	26	R[-1, 1]	190	12	15

4. Контрольные вопросы

1. Как можно классифицировать простые статистические гипотезы?
2. Проиллюстрируйте (опишите) область принятия и отклонения гипотезы при двустороннем критерии.
3. Дайте определение ошибки первого рода при проверке гипотезы.
4. Дайте определение ошибки второго рода при проверке гипотезы.
5. Нужно ли задавать уровень значимости при проверке нулевой (основной) и конкурирующей (альтернативной) гипотезы?
6. Как изменится ошибка второго рода при уменьшении/увеличении ошибки первого рода?
7. Как изменится область принятия и отклонения гипотезы при увеличении / уменьшении уровня значимости?
8. Как изменятся области принятия и отклонения гипотезы при одностороннем критерии, по сравнению с двусторонним критерием, при одном и том же уровне значимости?

Лабораторная работа № 3.

Решение задачи линейного корреляционного и регрессионного анализа

Цель работы – выявить связь между случайными переменными путем оценки коэффициентов корреляции и при установлении этой связи конкретизировать ее, построив регрессионную модель.

1. Теоретический обзор

Коэффициент корреляции ρ_{xy} характеризует тесноту связи между случайными переменными X и Y в генеральной совокупности. Коэффициент корреляции определяется через корреляционный момент (ковариацию) K_{xy} по формуле:

$$\rho_{xy} = \frac{K_{xy}}{\sigma_x \cdot \sigma_y} = \frac{M[(X-m_x)\cdot(Y-m_y)}{\sigma_x \cdot \sigma_y}. \quad (3.1)$$

Известно, что ρ_{xy} является показателем тесноты связи лишь в случае линейной зависимости между двумя переменными. Для линейно независимых случайных величин $\rho_{xy} \equiv 0$. Но даже и для зависимых СВ ρ_{xy} может быть равен 0. В этом случае СВ X и Y называют *некоррелированными*.

Пусть получена выборка N пар СВ X и Y. Тогда коэффициент корреляции можно оценить по выборочным данным следующим образом:

$$r = \rho_{xy} = \frac{S_{xy}}{S_x S_y}. \quad (3.2)$$

Вспомним "хорошие" (несмещённые, состоятельные и эффективные) оценки:

$$\bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i; \qquad \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i; \quad (3.3)$$

$$S_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} x_i^2 - N \cdot \bar{x}^2\right]; \quad (3.4)$$

$$S_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2 = \frac{1}{N-1}\left[\sum_{i=1}^{N} y_i^2 - N \cdot \bar{y}^2\right]; \quad (3.5)$$

$$S_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y}) = \frac{1}{N-1}\left[\sum_{i=1}^{N} x_i y_i - N \cdot \bar{x}\bar{y}\right]. \quad (3.6)$$

Тогда ***эмпирический коэффициент корреляции*** определяется по формуле:

$$r_{xy} = \frac{\frac{1}{N-1}\left[\sum_{i=1}^{N} x_i y_i - N\cdot\bar{x}\bar{y}\right]}{S_x \cdot S_y} = \frac{\sum_{i=1}^{N} x_i y_i - N\cdot\bar{x}\bar{y}}{\sqrt{\left(\sum_{i=1}^{N} x_i^2 - N\bar{x}^2\right)\left(\sum_{i=1}^{N} y_i^2 - N\bar{y}^2\right)}}. \quad (3.7)$$

Как и ρ_{xy} выборочный коэффициент корреляции принимает значения в интервале $[-1;+1]$, причем граничные значения достигаются только при

наличии идеальной линейной связи между наблюдениями. Нелинейная связь и (или) разброс данных, обусловленных неполной коррелированностью СВ или ошибками измерений, приводит к уменьшению абсолютного значения r_{xy}. Эмпирический коэффициент корреляции r_{xy} дает состоятельную, но смещённую оценку. Однако при $N > 50$ величина смещения составляет менее 1%. Для оценки точности выборочного значения r_{xy} удобно использовать некоторую функцию от r_{xy}:

$$W = \frac{1}{2}\ln\left[\frac{1+r_{xy}}{1-r_{xy}}\right]. \tag{3.8}$$

Распределение случайной величины W можно аппроксимировать нормальным распределением с соответствующим средним и дисперсией:

$$m_W = \frac{1}{2}\ln\left[\frac{1+\rho_{xy}}{1-\rho_{xy}}\right]; \quad \sigma_W^2 = \frac{1}{N-3}. \tag{3.9}$$

Даже для независимых случайных величин (СВ) эмпирический коэффициент корреляции может быть отличен от "0" вследствие случайного рассеивания результатов измерения. Т.е. из-за выборочной изменчивости необходимо проверять, свидетельствует ли не нулевые значения выборочного коэффициента корреляции о существовании статистически значимой корреляции между исследуемыми СВ X и Y. Сделать это можно, проверив гипотезу H_0: $\rho_{xy} = 0$, причем отклонение гипотезы будет свидетельствовать о принятии альтернативной гипотезы H_1 *−корреляция значимая*.

Из (3.9) следует, что при $\rho_{xy} = 0$ выборочное распределение W будет нормальным со средним $m_w = 0$ и дисперсией $\sigma_W^2 = \frac{1}{N-3}$. Поэтому область принятия гипотезы о нулевой корреляции будет иметь вид:

$$z_{\alpha/2} \le \frac{\sqrt{N-3}}{2}\ln\left[\frac{1+r_{xy}}{1-r_{xy}}\right] \le z_{1-\alpha/2}. \tag{3.10}$$

Здесь α −уровень значимости, z −стандартное нормальное распределение $N(0,1)$.

Если корреляционный анализ установит степень взаимосвязи двух и более случайных величин, логичен следующий шаг – построение модели этой связи. Такая модель дала бы возможность предсказать значения одной случайной величины по конкретным значениям другой. А методы решения подобных задач носят название "***регрессионный анализ***".

В линейный регрессионный анализ [3] входит широкий круг задач, связанных с построением (восстановлением) зависимостей между группами числовых переменных $X = (x_1, \dots x_p)$ и $Y = (y_1, \dots y_m)$. Предполагается, что X − независимые переменные (факторы, объясняющие переменные) влияют на значения Y − зависимых переменных (откликов, объясняемых переменных). По имеющимся эмпирическим данным (X_i , Y_i), i = 1, ..., n

требуется построить функцию $f(X)$, которая приближенно описывала бы изменение Y при изменении X:

$$Y \approx f(X).$$

Предполагается, что множество допустимых функций, из которого подбирается $f(X)$, является параметрическим:

$$f(X) = f(X, \Theta).$$

Здесь $\Theta -$ неизвестный параметр (вообще говоря, многомерный). При построении $f(X)$ будем считать, что

$$Y = f(X, \Theta) + \varepsilon, \tag{3.11}$$

где первое слагаемое – закономерное изменение Y от X, а второе – ε - случайная составляющая с нулевым средним; $f(X, \Theta)$ является условным математическим ожиданием Y при условии известного X и называется ***регрессией Y по X.***

Пусть X и Y одномерные величины; обозначим их x и y, а функция $f(x, \theta)$ имеет вид $f\ (x, \theta) = A + bx$, где $\theta = (A\ b)$. Учитывая имеющиеся наблюдения $(x_i, y_i), i = 1, \cdots, n$, полагаем:

$$y_i = A + bx_i + \varepsilon_i, \tag{3.12}$$

где $\varepsilon_i, \cdots, \varepsilon_n -$ независимые (ненаблюдаемые) одинаково распределенные случайные величины. Можно различными методами подбирать “лучшую” прямую линию. Общепринята такая процедура определения коэффициентов a и b, при которой минимизируется сумма квадратов отклонений наблюденных значений от предсказанных значений. Эта процедура называется ***методом наименьших квадратов*** **(МНК)**.

Построим оценку параметра $\theta = (A, b)$ так, чтобы величины

$$e_i = y_i - f(x_i, \theta) = y_i - A - bx_i,$$

называемые остатками, были как можно меньше, а именно, чтобы сумма их квадратов была минимальной:

$$\sum_{i=1}^{n} e_i^2 = \sum_{i}^{n} (y_i - A - bx_i)^2 = \min \text{ по } (A, b) \tag{3.13}$$

Чтобы упростить формулы, положим в (3.12) $x_i = x_i - \bar{x} + \bar{x}$, тогда получим:

$$y_i = a + b(x_i - \bar{x}) + \varepsilon_i, \quad i = 1, \cdots, n. \tag{3.14}$$

Здесь $\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad a = A + b\bar{x}$. Минимизируем сумму квадратов отклонений

$$Q = \sum_{i=1}^{N} (y_i - a - b(x_i - \bar{x}))^2, \tag{3.15}$$

приравняв нулю частные производные по a и b

$$\frac{\partial Q}{\partial a} = \frac{\partial Q}{\partial B} = 0. \tag{3.16}$$

Полученную систему линейных уравнений решим относительно a и b. Учитывая, что на практике у нас имеется ограниченная выборка из n пар наблюденных значений x и y, решение системы $(\hat{a}, \hat{b})$ легко находится:

$$\hat{a} = \bar{y}, \text{ где } \bar{y} = \sum_{i=1}^{n} y_i / n \tag{3.17}$$

$$\hat{b} = \frac{\sum_{i=1}^{n}(y_i-\bar{y})(x_i-\bar{x})}{\sum_{i=1}^{n}(x_i-\bar{x})^2}. \qquad (3.18)$$

Свойства оценок

Нетрудно показать, что если $M[\varepsilon] = 0,\ D[\varepsilon] = \sigma^2$, то

1) $M[\hat{a}] = a,\ \ M[\hat{b}] = b$, т.е. оценки несмещенные;
2) $D[\hat{a}] = \sigma^2/n,\ \ D[\hat{b}] = \sigma^2/\sum_{i=1}^{n}(x_i - \bar{x})^2$
3) $cov(\hat{a}, \hat{b}) = 0$;

Если дополнительно предположить нормальность распределения ε_i, то

4) оценки $\hat{a}$ и $\hat{b}$ нормально распределены и независимы;
5) остаточная сумма квадратов (3.15) независима от $(\hat{a}, \hat{b})$, а величина Q/σ^2 распределена по закону "хи-квадрат" χ^2_{n-2} с $n-2$ степенями свободы.

Оценка для σ^2 и интервальные оценки коэффициентов линейной регрессии

Свойство 5) дает возможность несмещенной оценки неизвестного значения σ^2 величиной

$$S^2 = Q/(n-2). \qquad (3.19)$$

Поскольку S^2 независима от $\hat{a}$ и $\hat{b}$, отношения

$$\sqrt{n}\cdot\frac{\hat{a}-a}{S} \quad \text{и} \quad \frac{\hat{b}-b}{S_b}, \text{ где } S_b = S/\sqrt{\sum_{i=1}^{n}(x_i-\bar{x})^2} \qquad (3.20)$$

имеют распределение Стьюдента с $(n-2)$ степенями свободы. Тогда соответствующие доверительные интервалы (при доверительной вероятности β) будут равны

$$\bar{a} - t_{\frac{\alpha}{2},n-2}\frac{S}{\sqrt{n}} < a < \bar{a} + t_{1-\frac{\alpha}{2},n-2}\frac{S}{\sqrt{n}},$$
$$\bar{b} - t_{\frac{\alpha}{2},n-2}S_b < b < \bar{b} + t_{1-\frac{\alpha}{2},n-2}S_b. \qquad (3.21)$$

Здесь $t_{\frac{\alpha}{2},n-2}$ и $t_{1-\frac{\alpha}{2},n-2}$ соответствующие квантили распределения Стьюдента с $n-2$ степенями свободы. Таким образом, найденные интервалы (3.21) с доверительной вероятностью $1-\alpha$ накрывают определяемые параметры (теоретические коэффициенты регрессии).

Проверка гипотез относительно коэффициентов линейной регрессии

На первом этапе регрессионного анализа наиболее важной является задача установления линейной зависимости между переменными y и x. С этой целью сформулируем гипотезы:

H_0: $\quad b = 0, -$ линейная зависимость отсутствует, коэффициент угла наклона прямой незначимо отличается от нуля;

H_1: $b \neq 0, -$ линейная зависимость значительная и коэффициент угла наклона не равен нулю.

При проверке гипотезы воспользуемся $t -$ статистикой и, если выполняется условие

$$t = \frac{b}{S_b} > t_{1-\frac{\alpha}{2}, n-2}, \tag{3.22}$$

то гипотезу H_0 следует отклонить при уровне значимости $\alpha = 1 - \beta$.

Другой способ (в данном случае эквивалентный (3.22)) проверки гипотезы H_0 состоит в вычислении статистики

$$F = \frac{\hat{b}^2 / D[\hat{b}]}{Q/\sigma^2(n-2)} = \frac{\hat{b}^2}{S_b^2} \tag{3.23}$$

распределенной, если H_0 верна, по закону $F(1, n-2)$ Фишера с числом степеней свободы 1 и $n-2$. Если

$$F > F(1-\alpha, 1, n-2), \tag{3.24}$$

где $F(1-\alpha, 1, n-2)$ –квантиль уровня 1 – α, то гипотеза H_0 отклоняется с уровнем значимости α.

Аналогичным образом проверяется гипотеза о статистической значимости нулю коэффициента регрессии a (свободный член линейного уравнения равен нулю): $t = \frac{a}{S}$.

Особый интерес представляет выборочное распределение $\hat{y}$ при конкретном значении $x = x_0$. Так как $\hat{y}$ ведет себя как СВ, распределенная по нормальному закону, для нее тоже можно построить доверительный интервал. Соответствующая статистика имеет вид:

$$t_{\tilde{y}} = \frac{\hat{y} - y}{S_{y|x}\left(\frac{1}{n} + \frac{(x_0 - \bar{x})^2}{\sum(x_0 - \bar{x})^2}\right)^{\frac{1}{2}}}. \tag{3.25}$$

В выражении (3.25) величина $S_{y|x} -$ это выборочное стандартное отклонение наблюденного значения y_i от предсказанного $\hat{y} = a + b(x_i - \bar{x})$, равное

$$S_{y|x} = \left[\frac{\sum(y_i - \hat{y}_i)^2}{n-2}\right]^{\frac{1}{2}} = \left[\frac{n-1}{n-2} \cdot S_y^2 (1 - r_{xy}^2)\right]^{\frac{1}{2}}. \tag{3.26}$$

Проверка качества уравнения регрессии

Оценим, насколько хорошо модель линейной регрессии описывает данную систему наблюдений. В качестве этой оценки воспользуемся <u>коэффициентом детерминации</u>.

Рассмотрим следующие вариации (суммы квадратов отклонений):

$T_{ss} = \sum_{i=1}^{n}(y_i - \bar{y})^2 -$(*total sum of square*) разброс фактических значений от их среднего арифметического;

$R_{ss} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2$ –(*regression sum of square*) разброс обусловленный регрессией от их среднего арифметического;

$E_{ss} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$ –(*error sum of square*) разброс за счет случайных отклонений от функции регрессии.

Оказывается,

$$T_{SS} = R_{SS} + E_{SS}, \tag{3.27}$$

т.е. полный разброс равен сумме разбросов за счет регрессии и за счет случайных отклонений. Величина R_{SS}/T_{SS} – это доля вариации значений y_i, обусловленной регрессией (т.е. доля закономерной изменчивости в общей изменчивости). Статистика

$$R^2 = R_{SS}/T_{SS} = 1 - E_{SS}/T_{SS} \tag{3.28}$$

называется ***коэффициентом детерминации***.

При $R^2 = 0$ регрессия ничего не дает, т.е. знание x не улучшает предсказания для y по сравнению с тривиальным $\hat{y}_i = \bar{y}$. Другой крайний случай $R^2 = 1$ означает точную подгонку: все точки наблюдений лежат на регрессионной прямой. Чем ближе к 1 значение R^2, тем лучше качество подгонки (регрессионной модели).

2. Корреляционный и регрессионный анализ в пакете Statistica 6.0

Анализ проведем с данными, представленными в файле **Product. sta.** Приведенные наблюдения по 45 предприятиям легкой промышленности, отражают статистические связи между стоимостью основных фондов (*Fonds,* млн руб.) и средней выработкой на 1 работника (*Product,* тыс. руб.). Также представлен вспомогательный признак – z: z = 1 – предприятие федерального подчинения, z = 2 – муниципальное.

Для построения корреляционной матрицы воспользуемся модулем: *Statistics → Basic Statistic and Tables → Correlation matrices*. Выберем все переменные и нажмем кнопку Summary.

Матрица коэффициентов корреляции – симметрична относительно главной диагонали. Значения на главной диагонали равны 1 и не указывают на строгую линейную зависимость, т.к. они получены при делении i – й дисперсии на саму себя. Остальные значения могут быть либо близки к нулю (менее 0.1), либо отличаются от 0. Возникает вопрос, эти флуктуации обусловлены статистикой выборки, либо переменные действительно коррелированны? Эту задачу можно корректно решить в рамках проверки гипотезы о равенстве нулю эмпирического коэффициента корреляции (гипотеза H_0), если коэффициент корреляции значимо отличается от нуля, то принимается альтернативная гипотеза – переменные коррелированны.

В приведенной таблице коэффициенты, значимо отличные от ноля выделены **красным**. Все оценки проведены для уровня значимости $\alpha = 0.05$.

Для получения более подробной информации в закладке *Options* отметим пункт Display r, p-levels.

Теперь матрица коэффициентов примет вид:

Variable	Correlations (Product.sta Marked correlations are N=45 (Casewise deletio		
	Fonds	Product	Z
Fonds	1,000(	,772З	-,1371
	p= ---	p=,00(	p=,369
Product	,772З	1,000(	-,2084
	p=,00(	p= ---	p=,17(
Z	-,1371	-,2084	1,000(
	p=,369	p=,17(	p= ---

Рис. 3.1. Результаты линейного корреляционного анализа

Анализ результатов свидетельствует, что для переменных Fonds и Product коэффициент корреляции незначимо отличается от 0 с вероятностью ~ 0.0001, т.е. принимаем альтернативную гипотезу – переменные коррелируют.

Убедимся, что предположение о линейной зависимости переменных не лишено смысла, для чего предварительно построим диаграмму рассеяния, выполнив последовательно действия *Graphs - 2D Graphs - Scatter plots - Variables - X: Fonds, Y : Product, Advanced, Graphs Type: Regular, Fit* (подбор): *Linear , Elipse, Normal coefficient 0.95- OK - OK* (см. рис. 3.2).

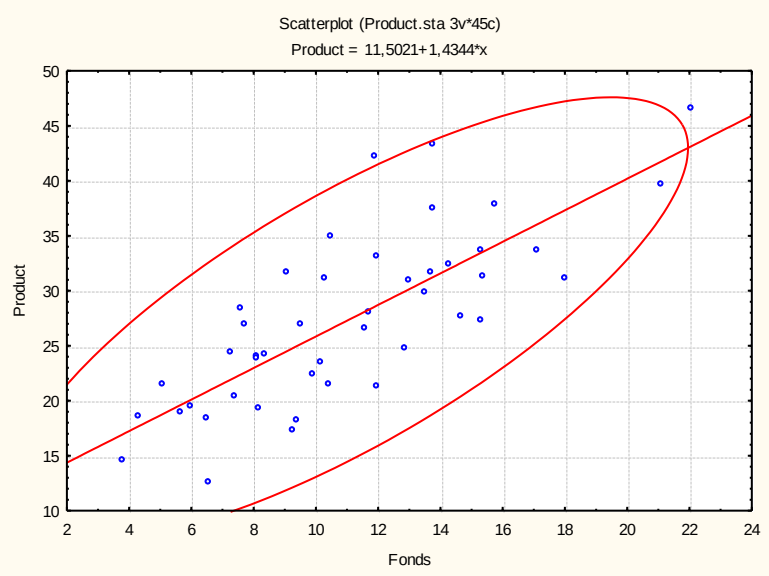


Рис. 3.2. Диаграмма рассеяния с подобранной прямой регрессии

Диаграмма рассеяния с наблюдениями, вытянутыми вдоль линии регрессии подтверждает наши предположения. На следующем этапе приступим к количественному анализу, используя при этом модуль ***Multiple Regression*** (множественная регрессия).

В стартовом диалоговом окне этого модуля при помощи кнопки *Variables* указываем зависимую переменную *Dependent var:* *Product* и независимую *Independent var:* *Fonds* - *OK*. В поле *Input file* указывается тип файла с данными *Raw Data* – данные в виде строчной таблицы. В поле *MD deletion* указываем способ исключения из обработки недостающих данных –

Casewise (игнорируется вся строка, в которой есть хотя бы одно пропущенное значение).

После выбора всех опций стартового диалогового окна регрессионного анализа и нажатия кнопки *OK* появляется окно результатов регрессионного анализа *Multiple Regressions Results*. Прежде, чем анализировать полученные результаты, опишем наиболее важные параметры полученной регрессионной модели:

- *Multiple R* – коэффициент множественной корреляции, характеризующий тесноту линейной связи между зависимой и всеми независимыми переменными;
- R^2 – коэффициент детерминации, выражающий долю вариации зависимой переменной, объясненную с помощью регрессионного уравнения;
- *adjusted R* – скорректированный коэффициент множественной корреляции. Включение новой переменной в регрессионное уравнение увеличивает R^2 не всегда, а только в том случае, когда частный *F*–критерий при проверке гипотезы о значимости включаемой переменной больше или равен 1. В противном случае включение новой переменной уменьшает значение R^2 и *adjusted R*;
- *F* – критерий используется для проверки значимости регрессии (в качестве нулевой гипотезы проверяется гипотеза – между зависимой и независимыми переменными нет линейной зависимости);
- *df* – числа степеней свободы для *F*– критерия;
- *p* – вероятность нулевой гипотезы для *F*– критерия;
- *Standard error of estimate* – стандартная ошибка оценки (уравнения); Эта оценка является мерой рассеяния наблюденных значений относительно регрессионной прямой;
- *Intercept* – оценка свободного члена уравнения;
- *Std.Error* – стандартная ошибка оценки свободного члена уравнения;
- *t* – критерий для оценки свободного члена уравнения;
- *p* – вероятность нулевой гипотезы для свободного члена уравнения.
- *Beta* – β – коэффициенты уравнения. Это стандартизированные регрессионные коэффициенты, рассчитанные по стандартизированным значениям переменных. По их величине можно оценить значимость зависимых переменных. Коэффициент показывает, на сколько единиц стандартного отклонения изменится зависимая переменная при изменении на одно стандартное отклонение независимой переменной, при условии постоянства остальных независимых переменных. Свободный член в таком уравнении равен 0.

В окне *Multiple Regression Results* получили такие результаты: коэффициент детерминации $R^2 = 0.597$; гипотеза о нулевом значении наклона отклоняется с высоким уровнем значимости $p = 0.000000$ (т.е. $p < 10^{-6}$).

Нажмем кнопку *Regression summary* – получим таблицу результатов (рис. 3.3).

N=45	Regression Summary for Dependent Variable: Product (Product.sta) R= ,77227708 R?= ,59641189 Adjusted R?= ,58702612 F(1,43)=63,544 p<,00000 Std.Error of estimate: 5,0082					
	Beta	**Std. Err. of Beta**	B	**Std. Err. of B**	**t(43)**	**p-Level**
Intercept			11,50212	**2,128204**	**5,404612**	**0,000003**
Fonds	0,772277	0,096880	1,43440	0,179942	7,971466	0,000000

Рис. 3.3. Результаты линейного регрессионного анализа

В ее заголовке повторены результаты предыдущего окна; в столбцах приведены: *B* – значения оценок неизвестных коэффициентов регрессии; *St. Err. of B* – стандартные ошибки оценки коэффициентов, *t* – значение статистики Стьюдента для проверки гипотезы о нулевом значении коэффициента; *p – level* – уровень значимости принятия этой гипотезы. В данном случае, поскольку значения *p-level* очень малы (меньше 10^{-5}), гипотезы о нулевых значениях коэффициентов отклоняются с высоким уровнем значимости. Итак, линейная модель имеет вид:

$$\textbf{\textit{Product}} = 11.5 + 1.43\ \textbf{\textit{Fonds}}.$$

Соответствующие стандартные ошибки коэффициентов равны: 2.1 и 0.18. Значение коэффициента детерминации R^2 = 0.597 достаточно велико (R = 0.77, т.е. 77 % всей изменчивости объясняется вариацией фондов).

Было бы логично предположить, что при более однородной совокупности предприятий – для предприятий федерального подчинения (z=1) регрессионная модель окажется более качественной. Предварительно визуально оценим данные процедурой *Scatterplot* (при отборе наблюдений используем кнопку *Select cases→Use selection conditions for this Analysis/Graph only→Include cases→Specific, selected by:→By Expression:* z=1. Сравнивая диаграммы рассеяния рис.3.2 и рис. 3.4 видим, что эллипс рассеяния более вытянут вдоль регрессионной прямой, причем все наблюдения находятся внутри эллипса.

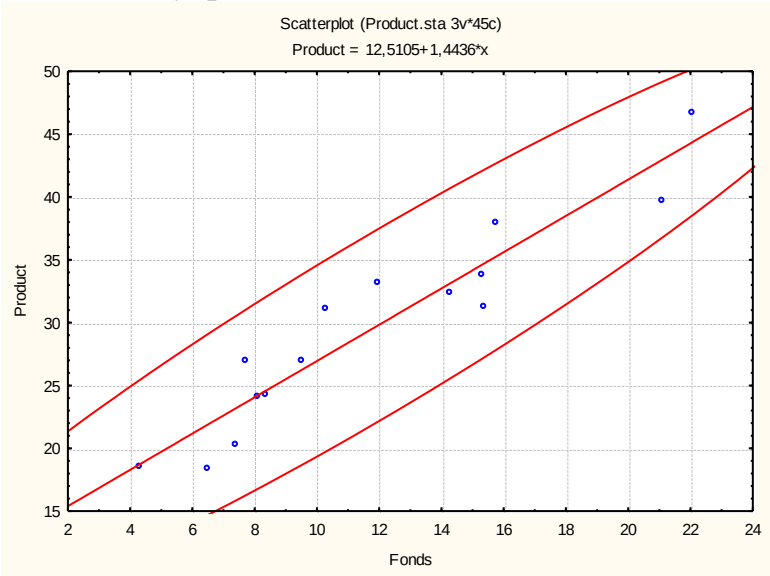


Рис. 3.4. Диаграмма рассеяния для предприятий федерального подчинения

Возвращаемся в окно *Multiple Regression* - *Select cases* - в окне *Case Selection Conditions* (условия выбора наблюдений *z* = 1) - *OK* - *OK* - в окнах *M.R.Results* и *Regression summary* получаем результаты:

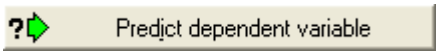
N=15	Regression Summary for Dependent Variable: Product (Product.sta) R= ,94717253 R?= ,89713581 Adjusted R?= ,88922318 F(1,13)=113,38 p<,00000 Std.Error of estimate: 2,6886					
	Beta	**Std. Err. of Beta**	B	**Std. Err. of B**	**t(43)**	**p-Level**
Intercept			12,51054	**1,753810**	**7,13335**	**0.000008**
Fonds	0,947173	0,088953	1,44356	0,135571	10,64802	0.000000

Рис. 3.5. Регрессионный анализ предприятий федерального подчинения

Теперь линейная модель примет вид:

$$\textbf{\textit{Product}} = 12.51 + 1.44\ \textbf{\textit{Fonds}}.$$

Коэффициент детерминации увеличился с 0.597 до 0.897, значения остальных параметров тоже улучшились (ошибки уменьшились).

Для расчета по полученному регрессионному уравнению значений зависимой переменной (***Product***) по значениям независимой переменной (***Fonds***) воспользуемся кнопкой 

(раздел Residuals/assumptions/prediction). Зададим значение ***Fonds*** = 18, и учтем, что в пакете Statistica приводится как точечная, так и интервальная оценка (см. рис. 3.6).

Variable	Predicting Values for (Product.sta) variable: Product Include condition: z=1		
	B-Weight	Value	B-Weight * Value
Fonds	1,443557	18,00000	25,98403
Intersept			12,51054
Predicted			38,49457
-95,0%CL			36,15750
+95,0%CL			40,83164

Рис. 3.6. Предсказанные точечные и интервальные оценки зависимой переменной

Анализ остатков

Остатки – это разности между опытными и предсказанными значениями зависимой переменной в построенной регрессионной модели.

Кнопка *Perform residual analysis* в модуле *Residuals/assumptions/prediction* запускает процедуру всестороннего анализа остатков регрессионного уравнения (см. рис. 3.7). При анализе остатков следует учитывать ряд существенных факторов:

- Если модель подобрана правильно, то остатки (столбец Residuals в *Predicted & Residuals Values*) будут вести себя достаточно хаотично, в известном смысле они будут напоминать белый шум.
- В остатках не будет систематической составляющей, резких выбросов, в чередовании их знаков не будет никаких закономерностей, остатки будут независимы друг от друга.

При анализе остатков весьма полезной характеристикой является расстояние Махаланобиса (Mahalanobis Distance). Независимые переменные в уравнении

регрессии можно представлять точками в многомерном пространстве (каждое наблюдение изображается точкой). В этом пространстве можно построить точку центра (среднюю точку). Эта "средняя точка" в многомерном пространстве называется центроидом, т.е. центром тяжести. Расстояние Махаланобиса определяется как расстояние от наблюдаемой точки до центра тяжести в многомерном пространстве. Соответственно, значения расстояния Махаланобиса, которые достаточно отличаются от остальных, указывают на выбросы.

	Predicted & Residual Values (Product.sta) Dependent variable: Product Include condition: z=1								
Case No.	Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val	Mahalanobis Distance	Deleted Residual	Cook's Distance
1	18,3000	21,8936	-3,5936	-1,0150	-1,3366	1,00691	1,03035	-4,1799	0,16952
2	31,1000	27,3791	3,7208	-0,2981	1,3839	0,72647	0,08886	4,0138	0,08137
3	27,0000	23,6259	3,3740	-0,7886	1,2549	0,89611	0,62197	3,7957	0,11071
4	37,9000	35,3187	2,5812	0,7396	0,9600	0,87425	0,54700	2,8864	0,06094
5	20,3000	23,1928	-2,8928	-0,8452	-1,0759	0,92237	0,71446	-3,2787	0,08752
6	32,4000	33,1534	-0,7534	0,4565	-0,2802	0,76780	0,20847	-0,8203	0,00379
7	31,2000	34,7413	-3,5413	0,6641	-1,3171	0,84238	0,44107	-3,9268	0,10471
8	39,7000	42,9696	-3,2696	1,7395	-1,2161	1,42978	3,02610	-4,5589	0,40659
9	46,6000	44,4131	2,1868	1,9282	0,8133	1,54970	3,71812	3,2749	0,24648
10	33,1000	29,8332	3,2667	0,0226	1,2150	0,69437	0,00051	3,5002	0,05653
11	26,9000	26,2243	0,6756	-0,4490	0,2513	0,76550	0,20164	0,7352	0,00303
12	24,0000	24,2033	-0,2033	-0,7131	-0,0756	0,86284	0,50863	-0,2267	0,00036
13	24,2000	24,6364	-0,4364	-0,6565	-0,1623	0,83932	0,43110	-0,4835	0,00157
14	33,7000	34,5969	-0,8969	0,6452	-0,3336	0,83478	0,41636	-0,9926	0,00657
15	18,5000	18,7178	-0,2178	-1,4301	-0,0810	1,24012	2,04531	-0,2767	0,00112
Minimum	18,3000	18,7178	-3,5936	-1,4301	-1,3366	0,69437	0,00051	-4,5589	0,00036
Maximum	46,6000	44,4131	3,7208	1,9282	1,3839	1,54970	3,71812	4,0138	0,40659
Mean	29,6600	29,6600	0,0000	0,0000	0,0000	0,95018	0,93333	-0,0358	0,08939

Рис. 3.7. Анализ остатков регрессионной модели

Для наглядного анализа поведения остатков построим их на нормальной вероятностной бумаге (*Normal plot of residuals*) рис. 3.8.

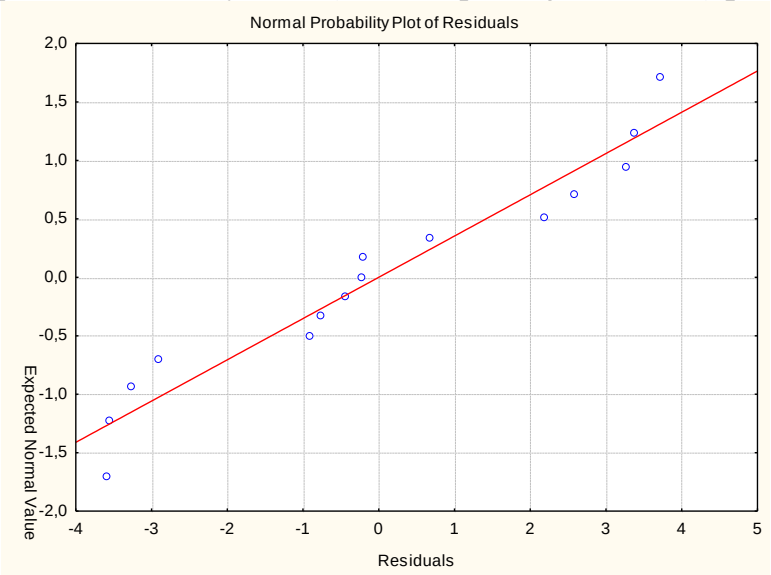


Рис. 3.8. Анализ остатков на нормальной вероятностной бумаге

Отсутствие больших отклонений и группирование остатков вдоль прямой свидетельствует о высоком качестве модели.

3. Задание

Необходимо исследовать зависимость урожайности y зерновых культур (ц/га) от ряда факторов (переменных) сельскохозяйственного производства [4], а именно:

x_1 - число тракторов на 100 га;

x_2 - число зерноуборочных комбайнов на 100 га;

x_3 - число орудий поверхностной обработки почвы на 100 га;

x_4 - количество удобрений, расходуемых на гектар (т/га);

x_5 - количество химических средств защиты растений, расходуемых на гектар (ц/га).

Исходные данные для 20 районов области приведены в табл. 3.1.

Таблица 3.1

	y	x_1	x_2	x_3	x_4	x_5
1	9.7	1.59	.26	2.05	.32	.14
2	8.4	.34	.28	.46	.59	.66
3	9.0	2.53	.31	2.46	.30	.31
4	9.9	4.63	.40	6.44	.43	.59
5	9.6	2.16	.26	2.16	.39	.16
6	8.6	2.16	.30	2.69	.32	.17
7	12.5	.68	.29	.73	.42	.23
8	7.6	.35	.26	.42	.21	.08
9	6.9	.52	.24	.49	.20	.08
10	13.5	3.42	.31	3.02	1.37	.73
11	9.7	1.78	.30	3.19	.73	.17
12	10.7	2.40	.32	3.30	.25	.14
13	12.1	9.36	.40	11.51	.39	.38
14	9.7	1.72	.28	2.26	.82	.17
15	7.0	.59	.29	.60	.13	.35
16	7.2	.28	.26	.30	.09	.15
17	8.2	1.64	.29	1.44	.20	.08
18	8.4	.09	.22	.05	.43	.20
19	13.1	.08	.25	.03	.73	.20
20	8.7	1.36	.26	.17	.99	.42

1. Создайте таблицу $5v \times 20c$. В первый столбец занесите значения переменной y, в остальные столбцы занесите данные, соответствующие вашему варианту (см. табл. 3.2).
2. Постройте матрицу коэффициентов парных корреляции и проанализируйте полученные результаты.
3. Создайте одномерную модель линейной регрессии, выбирая в качестве аргумента (независимой переменной) переменные, находящиеся в столбцах 2 – 5. Оцените качество полученных моделей, обоснуйте выбор лучшей.

Таблица 3.2

N	*II*	*III*	*IV*	*V*
1	x_1	x_2	$(x_1)^2$	$(x_2)^2$
2	x_1	x_3	x_1/x_3	$(x_3)^2$
3	x_1	x_4	$x_1 \times x_4$	$(x_4)^2$
4	x_1	x_5	x_5/x_1	$\ln(x_5)$

5	x_2	x_3	$\exp(x_2)$	$\exp(x_3)$
6	x_2	x_4	$\ln(x_4)$	$x_2 \times x_4$
7	x_2	x_5	x_2+x_5	$y \times x_5$
8	x_3	x_4	x_3/x_4	$\exp(x_4)$
9	x_3	x_5	x_5-x_3	$\ln(x_3)$
10	x_4	x_5	$y \times x_4$	$(x_5)^2$
11	x_1	x_2	$\ln(x_1)$	$\ln(x_2)$
12	x_1	x_3	$(x_1)^2$	x_3/x_1
13	x_1	x_4	x_1/x_4	x_4/ x_1
14	x_1	x_5	$x_5 \times x_1$	$\exp(x_5)$
15	x_2	x_3	$\exp(x_2)$	$x_2 \times x_3$
16	x_2	x_4	$\ln(x_2)$	x_2/x_4
17	x_2	x_5	x_2-x_5	$y \times x_5$
18	x_3	x_4	x_4/x_3	$\exp(y \times x_4)$
19	x_3	x_5	$\ln(x_5-x_3)$	$y \times x_3$
20	x_4	x_5	$y \times x_5$	$\ln(x_5)^2$
21	x_1	x_2	$\exp(x_1)$	$\exp(x_2)$
22	x_1	x_3	x_3/x_1	$\ln(x_3)$
23	x_1	x_4	$\ln(x_1 \times x_4)$	$\exp(x_4)$
24	x_1	x_5	$(x_5)^2$	$\exp(x_1)$
25	x_2	x_3	$y \times x_2$	$\ln(x_2 \times x_3)$
26	x_2	x_4	$\ln(x_2 \times x_4)$	x_2+x_4

4. Контрольные вопросы

1. С какой целью в линейном регрессионном анализе применяется метод наименьших квадратов?
2. Как изменится доверительный интервал для выборочного значения $\hat{y}$ при различных значениях аргумента x?
3. Что такое коэффициент детерминации?
4. Опишите процедуру проверки гипотез относительно коэффициентов линейной регрессии. Кстати, какое распределение имеет используемая при этом статистика?
5. Как проверить значимость (качество) уравнения регрессии? Какая статистика используется при этом?
6. Проанализируйте остатки в регрессионной модели. Каким требования (остаткам) отвечает качественная регрессионная модель?
7. Какой вид графика остатков на нормальной вероятностной бумаге подтвердит качество регрессионной модели?

Список литературы

1 Руководство пользователя пакета Statistica 5.1: http://exponenta.ru/soft/Statist/stat5_1/1/1.asp

2 Кобзарь А.И. Прикладная математическая статистика. Для инженеров и научных работников.- М.: ФИЗМАТЛИТ, 2006,-816 с.

3 Горицкий Ю.А. Практикум по статистике с пакетом Statistica (часть 2): http://exponenta.ru/educat/systemat/goritskii/part2/lr.asp

4 Магнус Я.Р., Катышев П.К., Пересецкий А.А. Эконометрика. Начальный курс. М.: Дело, 1998. 248с.

Кацман Юлий Янович

СТАТИСТИЧЕСКАЯ ОБРАБОТКА ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

Методические указания к лабораторным работам
(Цикл лабораторных работ)

Редактор
О.Н. Свинцова

Подписано к печати
Формат 60x84/16. Бумага офсетная.
Печать RISO. Усл. печ. л. 3,12. Уч.-изд. л. 2,98.
Тираж 120 экз. Заказ . Цена свободная.
Издательство ТПУ. 634050, Томск, пр. Ленина, 30.