Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования
Санкт-Петербургский горный университет

Кафедра информатики и компьютерных технологий

ЭКОНОМИКО-МАТЕМАТИЧЕСКИЕ МЕТОДЫ И МОДЕЛИРОВАНИЕ

ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Методические указания к лабораторным работам для студентов бакалавриата направления 21.03.02

САНКТ-ПЕТЕРБУРГ
2020

УДК 519.25 (073)

ЭКОНОМИКО-МАТЕМАТИЧЕСКИЕ МЕТОДЫ И МОДЕЛИРОВАНИЕ. Описательная статистика: Методические указания к лабораторным работам / Санкт-Петербургский горный университет.. Сост.: *В.В. Беляев, Т.Р. Косовцева.* СПб, 2020. 35 с.

Методические указания содержат теоретические сведения по выполнению лабораторной работы по дисциплине Экономико-математические методы и моделирование. Приведены необходимые теоретические сведения и примеры выполнения заданий по некоторым разделам теории вероятностей и математической статистики, которые являются теоретической основой применения корреляционно-регрессионного анализа. Все решения выполнены с использованием электронных таблиц MS Excel, в том числе с применением надстройки «Пакет анализа».

Предназначены для студентов бакалавриата направления 21.03.02 «Землеустройство и кадастры» (профиль «Городской кадастр»)

Научный редактор доц. *А.Б. Маховиков*

Рецензент канд. техн. наук *К.В. Столяров* (Телум Инк)

ВВЕДЕНИЕ

Корреляционно-регрессионный анализ является из самых применяемых математическим методом при решении задач, возникающих при рассмотрении проблем городского кадастра. Например, такой проблемой является массовая оценка объектов. Одним из краеугольных камней, лежащих в основе корреляционно-регрессионного анализа, является математическая статистика. *Математическая статистика* является частью статистики (от лат. *status* - состояние) - науке, изучающей, обрабатывающей и анализирующей количественные данные о самых разнообразных массовых явлениях окружающей нас жизни.

Основные задачи математической статистики - оценка неизвестных параметров распределений и проверка статистических гипотез.

На первом этапе с помощью массового наблюдения получают первичную информацию об отдельных фактах (единицах) изучаемого явления.

Собранная в ходе массового наблюдения информация представляет собой исходный материал для статистического исследования, для получения объективных выводов об изучаемом явлении.

Для того чтобы освободиться от влияния случайных причин и установить характерные черты изучаемого объекта, нужно получить сведения о достаточно большом числе единиц.

Следующим этапом статистического исследования является первичная обработка статистической информации: представление ее в виде удобно читаемых таблиц, изображение на диаграммах и вычисление наиболее показательных числовых характеристик. Методы, используемые на этом этапе, принято называть *описательной статистикой*. В настоящей работе описано применение этих методов с использованием электронных таблиц MS Excel.

ЛАБОРАТОРНАЯ РАБОТА

ТЕМА: ИЗУЧЕНИЕ БАЗОВЫХ ПОНЯТИЙ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Цель: Освоить на практике с помощью MS Excel построение вариационного и интервального рядов, вычисление выборочных характеристик (описательных статистик), получение статистических оценок и определение их свойств.

1 БАЗОВЫЕ ПОНЯТИЯ

При исследовании реальных экономических процессов приходится обрабатывать большие объёмы статистических данных, которые по своей сути являются случайными величинами (СВ). На практике количество реализаций СВ ограничено, что не позволяет применять напрямую теоретические методы анализа. Поэтому при обработке данных в первую очередь используют методы и модели математической статистики, позволяющие получить необходимые знания об исследуемом объекте.

1.1 ГЕНЕРАЛЬНАЯ СОВОКУПНОСТЬ И ВЫБОРКА

Статистической совокупностью называется множество предметов или явлений, объединённых в нечто целое и однородное по некоторым определенным признакам. Отдельные элементы, входящие в совокупность, называются *членами статистической совокупности*, а общее число членов совокупности – её *объёмом*.

Изменение признака при переходе от одного члена совокупности к другому называют его *вариацией*, а значение признака у отдельного члена статистической совокупности – его *вариантой*.

Выборочной совокупностью (или выборкой) называется совокупность случайно отобранных однородных элементов. *Генеральной совокупностью* называется совокупность всех однородных элементов, из которых произведена выборка.

Выборочная и генеральная совокупности, как правило, различаются объемами. Выборка называется *репрезентативной*, если она достаточно хорошо представляет исследуемый признак генеральной совокупности. Для обеспечения репрезентативности выборки применяют следующие способы отбора: *простой отбор*

(последовательно отбираются случайно попавшиеся объекты), *типический отбор* (объекты отбираются пропорционально представительству различных типов объектов в генеральной совокупности), *случайный отбор*, например, с помощью таблицы случайных чисел, и т.д.

Одной из основных задач статистического анализа является получение по заданной выборке достоверных сведений об интересующих исследователя свойствах и параметрах генеральной совокупности.

Основным типом значений переменных в математической статистике являются количественные переменные.

1.2 ВЫЧИСЛЕНИЕ ВЫБОРОЧНЫХ ХАРАКТЕРИСТИК

Значения количественных переменных являются числовыми, могут быть упорядочены и для них имеют смысл различные вычисления (например, вычисление среднего значения). На обработку количественных переменных ориентировано подавляющее большинство статистических методов.

Первый раздел математической статистики – описательная статистика – предназначен для представления исследуемых данных в удобном виде и для получения информации о них в терминах математической статистики и теории вероятностей. Для этого используются описательные или дескриптивные характеристики: минимум, максимум, размах, среднее, дисперсия, стандартное отклонение, медиана, квартили, мода.

При анализе конкретного показателя X все элементы выборки $x_1, x_2, ..., x_n$ объемом n обычно упорядочивают по неубыванию: $x_1 \le x_2 \le \ldots \le x_n$. Выборка, упорядоченная по неубыванию наблюдаемых значений, называется *вариационным рядом.* Разность между максимальным и минимальным значениями ряда X называется *размахом* выборки.

Если значение x_i встречается в выборке n_i раз, то число n_i называется *частотой (частостью)* значения x_i, а величина $m_i = \frac{n_i}{n}$ - *относительной частотой* значения x_i.

Пусть объем генеральной совокупности равен *N*. Тогда величина $\bar{x}_\Gamma = \frac{1}{N}\sum_{i=1}^{N} x_i$ является *генеральной средней*. *Генеральной дисперсией* является величина $D_\Gamma = \frac{1}{N}\sum_{i=1}^{N}\left(x_i - \bar{x}_\Gamma\right)^2$.

Генеральным средним квадратическим отклонением является величина $\sigma_\Gamma = \sqrt{D_\Gamma}$.

Так как реально чаще всего приходится работать с выборками из генеральной совокупности, то находят выборочные характеристики:

- выборочное среднее:

$$\bar{x}_в = \frac{1}{n}\sum_{i=1}^{n} x_i \qquad (1)$$

- выборочная дисперсия:

$$D_в = \frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_в\right)^2 \qquad (2)$$

- выборочное среднее квадратическое отклонение:

$$\sigma_в = \sqrt{D_в} = \sqrt{\overline{x^2} - \left(\bar{x}\right)^2} \qquad (3)$$

- выборочный коэффициент вариации *Vв*:

$$V_в = \frac{\sigma_в}{x_в}\cdot 100\% . \qquad (4)$$

Отметим, что приведенные выше формулы, аналогичны формулам для вычисления соответствующих характеристик дискретного распределения случайной величины в предположении, что все значения $x_1, x_2, ..., x_n$ равновероятны, т.е. вероятность появления каждого из них равна $\frac{1}{n}$.

В таблице 1 приведено соответствие между указанными выше характеристиками. Эта таблица может быть продолжена вполне очевидным образом как для параметров, приведенных ниже, так

других. Так моде генеральной совокупности $Mo_Г X$ соответствует $Mo_в X$. Для сокращения записи в дальнейшем индекс «*в*» будем опускать, т.е. для $Mo_в X$ будем использовать обозначение MoX .

Таблица 1

Параметр	*Генеральная совокупность*	*Выборочная совокупность.*	*Примечание*
	Оцениваемый параметр.	*Точечная оценка*	
Объем	N	n	
Среднее значение	$\overline{x}_Г$	$\overline{x}_в$	
Дисперсия	$D_Г$	$D_в$	Смещенная оценка
Дисперсия	$D_Г$	$D_{исп}$	Не смещенная оценка
Среднее квадратическое отклонение	$\sigma_Г$	$\sigma_{исп}$	

Мода MoX – это наиболее часто встречающееся значение признака в данном ряду распределения. Для дискретных вариационных рядов мода определяется как значение признака с наибольшей частотой. В случае непрерывной вариации мода может быть определена как значение признака, которому отвечает наибольшая плотность распределения частости.

Медианой называется такое значение варьирующего признака, которое делит ряд распределения на две равные части по объему частот. Медиана рассчитывается по-разному в дискретных и интервальных рядах.

Медианой $M_e X$ называется значение признака, относительно которого статистическая совокупность делится на две равные по объему части, причем в одной из них содержатся члены, у которых значения признака не больше, а в другой – члены со значениями признака не меньше, чем $M_e X$. Другими словами, медианой называется число, разделяющее выборку пополам: 50% элементов меньше медианы, а 50% -больше медианы.

Если в дискретном ряду распределения нечетное число уровней, то медианой будет серединное значение упорядоченного ряда признака, т.е. это элемент с номером $\frac{n+1}{2}$ вариационного ряда.

Если ряд распределения дискретный и состоит из четного числа членов, то медиана определяется как средняя величина из двух серединных значений вариационного ряда.

Квартили – это показатели, которые чаще всего используются для оценки распределения данных при описании свойств больших числовых выборок. В то время, как медиана разделяет упорядоченный массив пополам, квартили разбивают упорядоченный массив данных на четыре части.

Первый квартиль Q_1 – это число, разделяющее выборку на две части: 25% элементов меньше, а 75% - больше первого квартиля (5).

$$Q_1 = \frac{n+1}{4} \qquad (5)$$

Третий квартиль Q_3 – это число, разделяющее выборку на две части: 75% элементов меньше, а 25% - больше третьего квартиля(6).

$$Q_3 = \frac{3(n+1)}{4} \qquad (6)$$

Для вычисления квартилей применяются следующие правила.

1. Если индекс квартиля задается целым числом, значением квартиля считается элемент выборки с указанным индексом.
2. Если индекс квартиля задается величиной, представляющей собой среднее значение, вычисляемое по двум целым числам, квартиль равен среднему арифметическому, вычисленному по элементам, индексы которых равны эти двум числам.
3. Если индекс квартиля задается числом, которое не является целым и не кратно ½, он просто округляется до ближайшего целого. Квартилем является элемент выборки с указанным индексом.

Асимметрия – это свойство распределения выборки, которое характеризует несимметричность распределения СВ. На практике симметричные распределения встречаются редко, и чтобы выявить и оценить степень асимметрии, вводят следующую меру:

$$A_s = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^3}{\sigma^3}. \qquad (7)$$

Пределы значений A_s от $-\infty$ до $+\infty$ При $A_s = 0$ распределение симметрично: $MoX = \bar{x}$. При положительной асимметрии $MoX < \bar{x}$; при отрицательной - $MoX > \bar{x}$.

Эксцесс – это мера крутости кривой распределения. *Эксцесс* вычисляется по формуле:

$$E_k = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^4}{\sigma^4} - 3. \qquad (8)$$

Значения E_k лежат в открытом интервале $[-3,+\infty[$. Если $E_k > 0$, то кривая распределения имеет более острую вершину, чем нормальное распределение с параметрами $m = \bar{x}_в$ и $\sigma = \sqrt{D_в}$, и распределение называется островершинным. Если $E_k < 0$, то кривая распределения имеет более плоскую вершину, чем нормальное, и распределение называется плосковершинным.

Для нормального распределения $A_s = 0$, $E_k = 0$.

1.3 ЭМПИРИЧЕСКАЯ ФУНКЦИЯ РАСПРЕДЕЛЕНИЯ

Эмпирической функцией распределения называется следующая функция:

$$F(x) = \begin{cases} 0, & при\ x \le x_1 \\ \dfrac{i}{n}, & при\ x_i < x \le x_{i+1} \\ 1, & при\ x > x_n \end{cases}. \qquad (9)$$

Эта формула справедлива, когда все x_i различны.

1.4 ИНТЕРВАЛЬНЫЙ ВАРИАЦИОННЫЙ РЯД

Чтобы получить первое впечатление о распределении генеральной совокупности, необходимо провести некоторую обработку выборочных данных. Простейшей операцией является построение интервального ряда.

Если произвести группировку вариант по интервалам изменения признака (*интервальная группировка*) и результат представить рядом интервалов вариант, расположенных в порядке их возрастания, и рядом соответствующих частот, то получим *интервальный вариационный ряд*.

Под *частотой значения признака* или *интервала* понимают число членов совокупности, варианты которых лежат в данном интервале. Отношение частоты к объему совокупности называется относительной *частотой* или *частостью*.

Число равных интервалов k, на которые следует разбить весь диапазон значений признака $X\left[x_{min,}\ x_{max}\right]$, может быть найдено по формуле (10):

$$k = log_2 n + 1, \tag{10}$$

где n – объем статистической совокупности.

Число интервалов должно быть не меньше 8-10 и не больше 20-25.

Размах выборки определяется по формуле:

$$\Delta = x_{max} - x_{min}, \tag{11}$$

а длина интервала - по формуле:

$$h = \frac{\Delta}{k}. \tag{12}$$

Формулы (10) и (12) дают оценочное значение количества интервалов и их размеров. При практическом построении рекомендуется брать значения k и h, которые соответствуют здравому смыслу.

В дальнейшем будем использовать следующие обозначения: a_i, b_i - левая и правая границы i-го интервала соответственно; x_i – середина этого интервала; m_i – частота интервала.

1.5 ГРАФИЧЕСКОЕ ПРЕДСТАВЛЕНИЕ ИНТЕРВАЛЬНЫХ ВАРИАЦИОННЫХ РЯДОВ

Для наглядного представления статистического распределения пользуются графическим изображением интервальных вариационных рядов. К числу таких графических изображений относятся гистограмма, полигон, кумулята.

1). Построение гистограммы.

Для построения гистограммы нужно составить таблицу, в которой необходимо указать границы интервалов, найти их середины и частоту значений признака для каждого интервала.

Пример

Пусть объем выборки равен n=60; x_1=-2,18; x_{60}=12,04; h=2; k=9. В табл.2 представлен соответствующий интервальный вариационный ряд.

Таблица 2

№ (разряд)	Границы интервала		Середина интервала *x[i]*	Частота *mi*
	левая *a[i]*	правая *b[i]*		
1	-4	-2	-3,00	1
2	-2	0	-1,00	3
3	0	2	1,00	6
4	2	4	3,00	12
5	4	6	5,00	20
6	6	8	7,00	8
7	8	10	9,00	8
8	10	12	11,00	1
9	12	14	13,00	1
sum				60

По оси абсцисс откладывают интервалы значений признака, и на каждом из них, как на основании, строят прямоугольник с высотой, пропорциональной частоте интервала.

Гистограмма, построенная с помощью Мастера диаграмм программы MS Excel, приводится на рис.1.

Рис. 1. Гистограмма, построенная с помощью Мастера диаграмм MS Excel

2). Построение полигона

Для построения *полигона* на оси абсцисс откладывают интервалы значений признака, в серединах интервалов восстанавливают перпендикуляры, длины которых пропорциональны соответствующим частотам, затем концы соседних перпендикуляров соединяют отрезками прямых, а концы крайних перпендикуляров соединяют с серединами соседних интервалов, частоты которых равны нулю. В результате получим замкнутую фигуру в виде многоугольника.

Полигон для интервального ряда приведен на рис.2.

3). Построение кумуляты

Накопленной частостью (частотой) в точке x называют суммарную частость (частоту) членов статистической совокупности со значениями признака меньшими, чем x.

Если в вариационном ряду вместо частот или частостей записать соответственно накопленные частоты или частости, то получится *кумулятивный ряд*. Для графического построения кумулятивных рядов пользуются *кумулятами*.

Кумулята (рис.3) строится следующим образом: на оси абсцисс отмечают точки, соответствующие границам интервалов или значениям признака. В каждой такой точке восстанавливают перпендикуляр, длина которого пропорциональна накопленной

частоте. Концы соседних перпендикуляров соединяют отрезками. Полученная ломаная линия называется *кумулятой*.

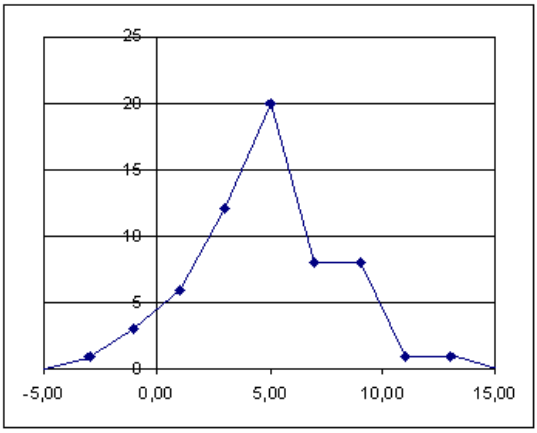


Рис. 2. Полигон интервального ряда

Эмпирическая функция распределения отличается от кумуляты только масштабом.

Рис. 3 Кумулята интервального ряда

Используя полученный интервальный ряд, можно вычислить все описательные характеристики, полагая, что все варианты выборки, лежащие внутри *i*-го интервала, принимают значения равные x_i с частотой m_i. Тогда выборочное среднее $\overline{x}_в$ и выборочная дисперсия $D_в$ вычисляются по формулам:

$$\overline{x}_в = \frac{1}{n}\sum_{i=1}^{k} m_i x_i \,, \tag{13}$$

$$D_в = \frac{1}{n}\sum_{i=1}^{k} m_i \left(x_i - \overline{x}_в\right)^2 . \tag{14}$$

Если $a_j - b_j$ - *модальный интервал*, т.е. интервал, которому соответствует наибольшая частота m_j, а интервалы вариационного ряда имеют постоянную ширину *h*, то мода признака вычисляется по формуле

$$MoX = a_j + h \cdot \frac{m_j - m_{j-1}}{\left(m_j - m_{j-1}\right) + \left(m_j - m_{j+1}\right)}, \tag{15}$$

где m_{j-1}, m_{j+1} – частоты интервалов, предшествующих модальному и следующего за модальным, соответственно.

Для интервального распределения сначала находят так называемый *медианный интервал* $a_s - b_s$, номер которого вычисляют из неравенств

$$\gamma\left(a_s\right) \le 0{,}5\,; \quad \gamma\left(b_s\right) > 0{,}5\,; \tag{16}$$

где $\gamma(x)$ - накопленная частота в точке *x*. В предположении, что в медианном интервале признак распределен равномерно, медиана признака *X* определяется по формуле:

$$MeX = a_s + h \cdot \frac{\frac{n}{2} - \gamma(a_s)}{m_s}, \tag{17}$$

где *h* – ширина интервала *c* номером *s*; m_s – частота этого интервала.

2 ДИАГРАММА ТИПА “ЯЩИК С УСАМИ”

2.1 ОБЩИЕ СВЕДЕНИЯ

Диаграмма типа “ящик с усами” изображает важные характеристики описательной статистики на одном компактном рисунке. Она предложена Джоном Тьюки (John Tukey) в 1977 г. в основополагающей книге *Exploratory Data Analysis*. Диаграмма типа “ящик с усами” отображает следующие характеристики СВ:

1. первый квартиль, медиана, третий квартиль и интерквантильный диапазон;
2. минимальное и максимальное значения;
3. умеренные и экстремальные выбросы.

Диаграмма типа “ящик с усами” дает хорошее визуальное представление изменчивости данных, а также асимметрии распределения. Типичный вид диаграммы типа “ящик с усами” приведен на рис.4

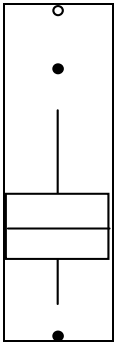


Рис. 4. Диаграмма типа “ящик с усами”

2.2. ИНТЕРКВАРТИЛЬ

Первый компонент диаграммы типа “ящик с усами” называется *интерквартиль* или *интерквартильный диапазон* (*interquartile range — IQR*), который простирается от первого до третьего квартиля.

Интерквартиль (*IQR)* - одна из мер разброса или рассеяния данных. Он равен разности между верхним и нижним (первым и третьим) квартилями. Другими словами *IQR* - это ширина интервала, содержащего средние 50% выборки. Таким образом, чем меньше *IQR*,

тем меньше рассеяние. Положительной чертой этого показателя является его устойчивость (робастность), т.е. на него слабо влияют выбросы.

Пример

Пусть дана выборка (уже в виде вариационного ряда):

2 3 4 5 6 6 6 7 7 8 9.

Ее верхний квартиль равен 7, ее нижний квартиль равен 4, следовательно, IQR равняется 7 - 4 = 3.

Для создания интерквартиля строят прямоугольник (“ящик”) от первого до третьего квартиля. Внутри ящика проводят горизонтальную линию на уровне медианы (второго квартиля) (рис.5).

2.3 ОГРАЖДЕНИЯ

После построения интерквартильного диапазона можно приступать к вычислению внутреннего и внешнего ограждений. *Внутренние ограждения* (*inner fences*) располагаются в области, большей третьего квартиля *плюс величина* $1{,}5 \times IQR$ или меньшей первого квартиля *минус величина* $1{,}5 \times IQR$. *Внешние ограждения* (*outer fences*) располагаются в области большей третьей квартили $+3 \times IQR$ или меньшей первой квартили $-3 \times IQR$ (рис.5).

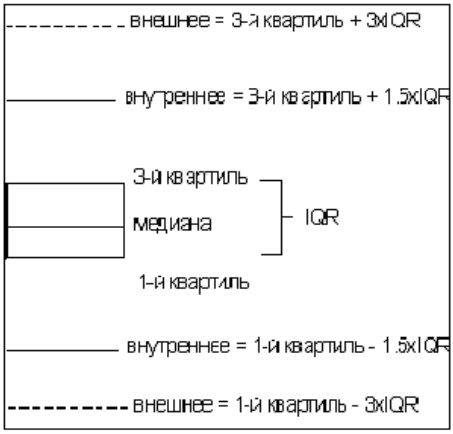


Рис. 5. Расположение ограждений при построении диаграммы «ящик с усами»

Замечание. *Диаграммы на рис.5-8 нарисованы без* точного соответствия масштабу.

2.3 ВЫБРОСЫ

Все значения выборки, которые лежат в промежутке между внутренним и внешним ограждениями, называются *умеренными выбросами* (*moderate outlier*) и обозначаются символами ●.

Все значения, которые лежат за пределами внешних ограждений, называются *экстремальными выбросами* (*extreme outlier*) и обозначаются символами ○ (рис.6).

Рис. 6. Выбросы при построении диаграммы «ящик с усами»

2.4. УСЫ

Усы - вертикальные линии, проведенные от «ящика» до максимального и минимального значения СВ внутри внутреннего ограждения (рис.7), такие значения *не* считаются выбросами.

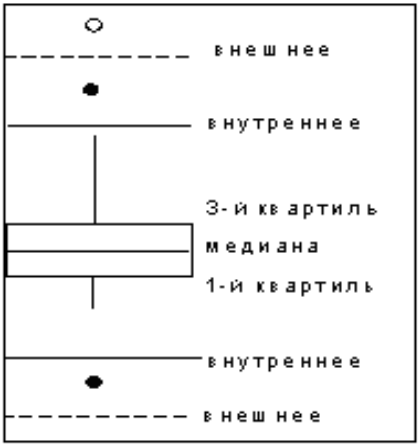


Рис. 7. Расположение усов при построении диаграммы «ящик с усами»

2.5 ОКОНЧАТЕЛЬНЫЙ ВИД ДИАГРАММЫ

Обычно в окончательном виде статистической диаграммы типа “ящик с усами” внутреннее и внешнее ограждения не отображаются. Обычно эта диаграмма выглядит так, как показано на рис.8. Как видите, в этих данных имеются три выброса, причем один из них является экстремальным, а распределение в целом асимметрично.

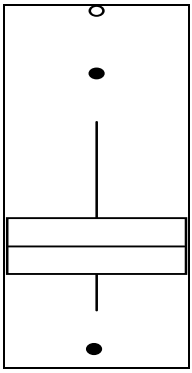


Рис. 8. Окончательный вид статистической диаграммы типа “ящик с усами”

2.6 “ЯЩИКИ С УСАМИ” И РАСПРЕДЕЛЕНИЯ

“Ящик с усами” дает уникальное представление данных и широко используется в представлении экономической и технической информации. На рис.9 приведен пример использования диаграммы типа “ящик с усами” для иллюстрации изменения количества процессоров, используемых в течение ряда лет в одной из сетей суперкомпьютеров [2].

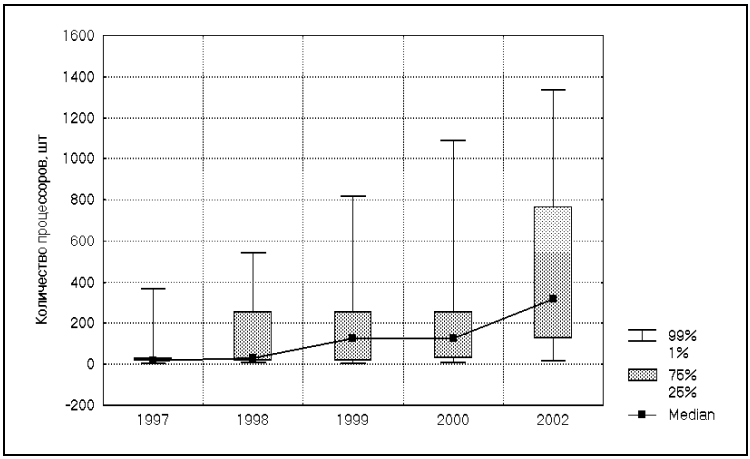


Рис. 9. Диаграмма размаха значений количества процессоров суперкомпьютеров вычислительных центров коллективного пользования сети DREN по годам

3 СТАТИСТИЧЕСКИЕ ВЫВОДЫ: ОЦЕНКИ И ПРОВЕРКА ГИПОТЕЗ

Статистические выводы - это заключения о свойствах генеральной совокупности, полученные на основе исследования выборки, случайно отобранной из генеральной совокупности. Например, анализируется доход (X) населения некоторого достаточно большого города. Этот анализ может быть осуществлен на основе выборки определенного объема (пусть n=1000).

Для выборочных данных определяем средний доход $\bar{x} = \dfrac{\sum x_i}{n}$ и разброс данных $s^2 = \sum \left(x_i - \bar{x}\right)^2 / n$. Далее возникает естественный вопрос: можно ли ожидать, что аналогичные значения будут такими же для всего города? И можно ли обобщить результаты,

полученные по выборке, на генеральную совокупность? В этом назначение статистических выводов.

На основе выборки можно получить лишь оценки параметров генеральной совокупности, так как эти оценки строятся на основе ограниченного набора данных; эти значения оценок могут изменяться от выборки к выборке. Процесс нахождения оценок параметров генеральной совокупности по определенному правилу называется оцениванием.

Выделяют два типа оценивания:

- оценивание *вида* распределения;
- оценивание *параметров* распределения.

В качестве оценки вида распределения можно взять выборочное распределение, а в качестве оценок параметров распределения генеральной совокупности используют выборочные оценки этих параметров.

Различают два вида оценок параметров – *точечные* и *интервальные*. Когда оценка определяется одним числом, то эта оценка является *точечной*.

Оценка является *интервальной*, если она выражается некоторым интервалом на числовой оси. Очевидно, что интервал может быть задан двумя числами – его концами.

После определения оценок обычно встает вопрос об их качестве и статистической значимости.

Пусть рассматривается генеральная совокупность наблюдаемой СВ X. Для оценки её параметра Θ из генеральной совокупности извлекается выборка объема n: $x_1, x_2, \ldots, x_n$.

На основе этой выборки может быть найдена оценка Θ^* параметра Θ.

Точечной оценкой Θ^* *параметра* Θ называется числовое значение этого параметра, полученное по выборке объема n.

Например, для нормального распределения $N(m, \sigma^2)$ параметрами являются математическое ожидание m и среднее квадратическое отклонение σ.

Точечными оценками m и σ могут быть значения $m^* = \bar{x}_в = \frac{1}{n}\sum_{i=1}^{n} x_i$ и $\sigma^* = \sigma_в = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(x_i - \bar{x}_в\right)^2}$, соответственно.

Очевидно, что оценка Θ^* является функцией от выборки, то есть $\Theta^*=\Theta^*(x_1,x_2,\ldots,x_n)$. А так как выборка носит случайный характер, то оценка Θ^* является случайной величиной, принимающей различные значения для различных выборок.

Любую оценку $\Theta^*=\Theta^*(x_1,x_2,\ldots,x_n)$ называют *статистической оценкой* параметра Θ.

Качество оценок характеризуется следующими основными свойствами: **несмещенность, эффективность и состоятельность.**

Оценка Θ^* называется *несмещенной* оценкой параметра Θ, если ее математическое ожидание равно оцениваемому параметру:

$$M(\Theta^*)=\Theta. \tag{18}$$

Оценка Θ^* называется *эффективной* оценкой параметра Θ, если ее дисперсия $D(\Theta^*)$ меньше дисперсии любой другой оценки, полученной по выборке объемом n.

Оценка параметра Θ называется *асимптотически эффективной,* если с увеличением объема выборки ее дисперсия стремится к нулю, то есть

$$D\left(\Theta_n^*\right) \to 0 \text{ при } \quad n \to \infty \tag{19}$$

Оценка Θ_n называется *состоятельной* оценкой параметра Θ, если Θ_n сходится по вероятности к Θ при $n\to\infty$, т.е. для любого $\varepsilon > 0$ при $n\to\infty$

$$\forall \varepsilon > 0 \qquad \lim_{n\to\infty} P\left(\left|\Theta_n^* - \Theta\right| < \varepsilon\right) = 1 \tag{20}$$

Иными словами, если оценка состоятельна, то событие, состоящее в том, что разница между истинным значением параметра и его оценкой сколь угодно мала при достаточно большом объеме выборки, становится достоверным.

Некоторые свойства выборочных оценок

Доказано [1], что выборочное среднее $\bar{x}_в = \frac{1}{n}\sum_{i=1}^{n} x_i$ является несмещенной и состоятельной оценкой математического ожидания $\bar{x}_Г = M(X)$ генеральной совокупности.

Выборочная дисперсия $D_в = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x}_в)^2$ является смещенной оценкой дисперсии генеральной совокупности $D_г = D(X) = \sigma^2$ и, как следствие, выборочная дисперсия оценивает генеральную дисперсию неточно.

Для таких случаев следует использовать *исправленную дисперсию*

$$S^2 = \frac{n}{n-1} D_в = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x}_в)^2 . \qquad (21)$$

Исправленная дисперсия S^2 является несмещенной и состоятельной оценкой дисперсии $D_г$ СВ X.

Необходимо отметить, что при n >30 различие между $D_в$ и S^2 практически незначимо.

Точечная оценка Θ^* (выражается одним числом) по данным выборки дает оценочное значение соответствующего параметра генеральной совокупности Θ. Такая оценка имеет два существенных недостатка:

– при малой выборке разница между значением параметра и оценкой может быть велика;

– не удается статистически оценить величину указанной разницы.

Чтобы оценить точность и достоверность оценки находят *интервальные оценки параметров.* Интервальная оценка определяется двумя числами - концами интервала.

Пусть для оценки параметра Θ генеральной совокупности используется выборка $x_1, x_2, \ldots, x_n$., Θ_L и Θ_U - такие значения, что выполняется равенство

$$P(\Theta_L < \Theta < \Theta_U) = \gamma ,$$

где Θ - оцениваемый параметр; γ - выбранная исследователем вероятность. Тогда случайный интервал $]\Theta_L,\Theta_U[$ называется *доверительным интервалом* для оценки параметра Θ с мерой надежности γ.

Другими словами: случайный интервал $]\Theta_L,\Theta_U[$ называется *доверительным интервалом* для оценки параметра Θ с *мерой надежности* γ, если с вероятностью γ он покрывает оцениваемый параметр. По смыслу определения надежности, величину γ следует брать близкой к единице. На практике величина γ обычно принимает значения 0,9; 0,95 или 0,99.

Длина интервала есть случайная величина, зависящая от выборки ($x_1, x_2, \ldots, x_n$). Случайно и положение интервала на числовой оси. Интервал обычно симметричен относительно точечной оценки параметра Θ^*.

В этом случае для некоторого ε выполняется соотношение $P(\Theta^*-\varepsilon< \Theta < \Theta^*+\varepsilon)= \gamma$, следовательно доверительный интервал можно представить в виде $]\Theta^*-\varepsilon,\ \Theta^*+\varepsilon[$, и для его построения необходимо определить ширину 2ε или полуширину - величину ε. Вид выражения (формулы) для вычисления полуширины доверительного интервала ε для оценки параметра Θ зависит от того, какая предварительная информация о распределении СВ известна.

Пусть случайная величина распределена нормально с параметрами m, σ^2, т.е. $\left(X \sim N(m, \sigma^2)\right)$.

Для оценки параметра m генеральной совокупности при неизвестном значении σ, в качестве точечной оценки m_x используют $\bar{x}_в$, а доверительный интервал имеет вид**:**

$$\bar{x}_в - \frac{t_{1-\gamma,n-1}\cdot S}{\sqrt{n}} < m < \bar{x}_в + \frac{t_{1-\gamma,n-1}\cdot S}{\sqrt{n}}, \qquad (22)$$

где S – исправленное среднее квадратическое отклонение случайной величины X, вычисленное по выборке ($x_1, x_2, \ldots, x_n$), $S=\sqrt{S^2}$.

Таким образом, полуширина равна

$$\varepsilon = \frac{t_{1-\gamma,n-1} \cdot S}{\sqrt{n}}, \qquad (23)$$

где $t_{1-\gamma,n-1}$ - критическое значение распределения Стьюдента с n-1 степенями свободы для p-значения, равного $(1-\gamma)$.

Учитывая свойства нормального распределения, для которого параметр m равен математическому ожиданию, формулу (22) можно переписать в виде

$$\bar{x}_{\text{в}} - \frac{t_{1-\gamma,n-1} \cdot S}{\sqrt{n}} < \bar{x}_{\Gamma} < \bar{x}_{\text{в}} + \frac{t_{1-\gamma,n-1} \cdot S}{\sqrt{n}} \ , \qquad (24)$$

где $\bar{x}_{\Gamma}$ - среднее генеральной совокупности.

4 НАДСТРОЙКА «ПАКЕТ АНАЛИЗА» MS EXCEL

Надстройка (модуль) «Пакет анализа» MS Excel предназначен для выполнения базовых операций статистического анализа. Полученные с его помощью результаты не обновляются при изменении исходных данных, поэтому после их изменения для обновления результатов требуется снова выполнить соответствующую команду.

Для активизации надстройки **Пакет Анализа** выполните команду **Параметры Excel** выберите пункт меню **Надстройки** →кнопка **Перейти**…→**Пакет Анализа.** Модуль доступен из пункта меню **Данные,** группа **Анализ.**

4.1 ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Это средство анализа (рис.10) служит для создания таблицы с точечными оценками одномерной выборки

Раздел *Входные данные*

Поле **Входной интервал** используется для ввода диапазона смежных ячеек с анализируемыми данными.

Группа переключателей **Группирование** используется для указания способа расположения анализируемых данных по столбцам или по строкам.

Флажок **Метки в первой строке** устанавливают для обозначения того, что первая строка анализируемых данных содержит заголовки столбцов.

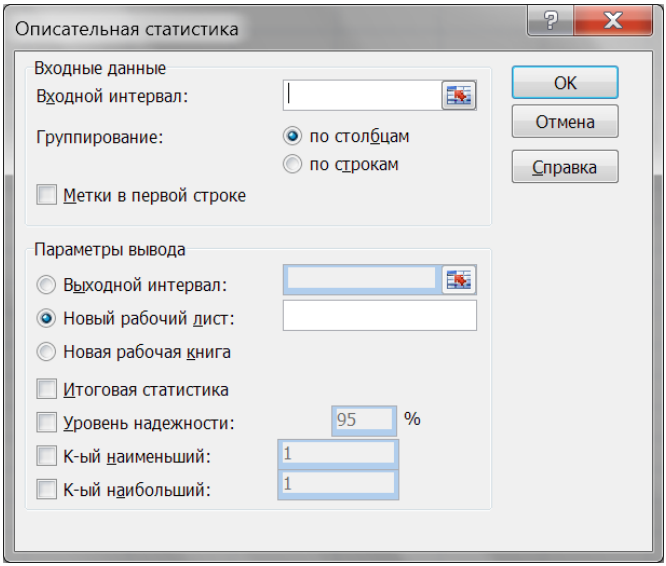


Рис. 10. Диалоговое окно «Описательная статистика»

Раздел *Параметры вывода*

Выходной интервал - переключатель, используемый для указания начальной ячейки в верхнем левом углу диапазона ячеек, в которых будут располагаться полученные результаты.

Переключатель **Новый рабочий лист** используется для указания того, что результаты будут располагаться на новом рабочем листе с указанным именем.

Флажок **Итоговая статистика** используется для вывода статистических параметров.

4.2 РАНГ И ПЕРСЕНТИЛЬ

Инструмент «Ранг и персентиль» - средство анализа, которое используется для вывода таблицы, содержащей порядковый и процентный ранги для каждого значения в наборе данных (рис.11).

Данная процедура может быть применена для анализа относительного взаиморасположения данных в наборе и для приближенного построения функции распределения.

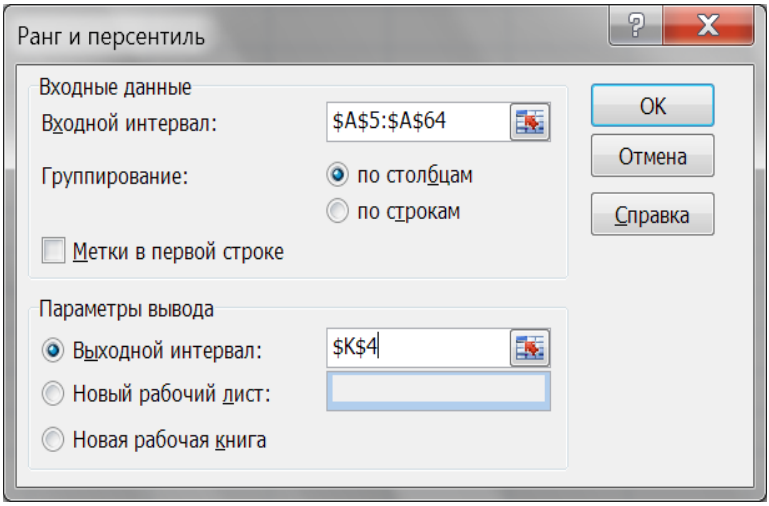


Рис. 11. Диалоговое окно «Ранг и персентиль»

4.3 ГИСТОГРАММА

«Гистограмма» - один из инструментов пакета анализа. Используется для вычисления выборочных и интегральных частот попадания данных в указанные интервалы значений (рис.12).

Параметры диалогового окна "Гистограмма":

Входной интервал. Используют для ввода диапазона смежных ячеек с исследуемыми данными.

Интервал карманов (необязательный). Используют для ввода диапазона ячеек и необязательного набора граничных значений, определяющих отрезки (карманы). Эти значения должны быть введены в возрастающем порядке.

В MS Excel вычисляется число попаданий данных в интервал, ограниченный началом отрезка и соседним большим по порядку, если такой есть. При этом в интервал включаются значения, принадлежащие нижней границе отрезка и не включаются значения, соответствующие верхней границе.

Если диапазон карманов не был введен, то набор отрезков, равномерно распределенных между минимальным и максимальным значениями данных, будет создан автоматически.

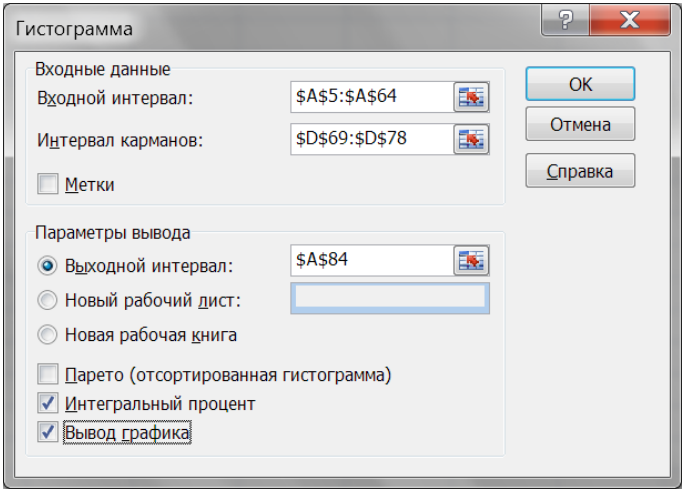


Рис. 12. Диалоговое окно «Гистограмма»

Метки. Флажок устанавливают, когда первая строка анализируемых данных содержит заголовки столбцов. Если заголовки отсутствуют; в этом случае подходящие названия для данных выходного диапазона будут созданы автоматически.

Выходной интервал. Используют для указания ссылки на левую верхнюю ячейку выходного диапазона. Размер выходного диапазона будет определен автоматически, и на экран будет выведено сообщение в случае возможного наложения выходного диапазона на исходные данные.

ЗАДАНИЕ

Из генеральной совокупности извлечена выборка объема n. Изучить распределение непрерывного признака X некоторой генеральной совокупности.

Требуется:

1. Построить вариационный ряд.
2. Найти точечные оценки математического ожидания, дисперсии, среднего квадратического отклонения, моды, медианы, размаха, асимметрии и эксцесса.

3. Вычислить первый, второй и третий квартили, используя встроенную функцию КВАРТИЛЬ();
4. Вычислить 5-ый, 50-ый и 95-ый персентили, используя встроенную функцию ПЕРСЕНТИЛЬ().
5. Построить (приближенно) эмпирическую функцию распределения данного вариационного ряда. С помощью графика этой функции проиллюстрировать результаты, полученные в п.3 и 4.
6. Построить диаграмму типа «Ящик с усами».
7. Построить интервальный вариационный ряд.
8. Построить полигон, гистограмму, кумуляту и эмпирическую функцию распределения для полученного интервального вариационного ряда.
9. Найти точечные оценки числовых характеристик m_X, D_X, σ_X используя интервальный ряд. Сравнить результаты с п.2 (объяснить различия).
10. Считать, что выборка получена из генеральной совокупности нормально распределенной $N(m, \sigma^2)$ с неизвестным σ. Построить доверительный интервал для оценки математического ожидания генеральной совокупности (параметра m). Вычисления провести по формулам (22)-(23). Используя последнюю формулу, проанализировать зависимость ширины доверительного интервала от объема выборки, однородности выборки и меры надежности γ.

<u>Указания по выполнению лабораторной работы</u>.

Исходные данные приведены в виде таблиц Excel, в интервале ячеек A5:A64 (рис.13).

Для упорядочивания признака X по возрастанию следует воспользоваться командой **Данные→Сортировка**, результат сортировки расположить в этом же интервале.

Для выполнения п.2. точечные оценки можно найти двумя способами:

- с помощью встроенных функций: СРЗНАЧ(), ДИСП(), МЕДИАНА() и т.д. (рис.13 и рис.14);
- с помощью надстройки MS Excel *«Пакет Анализа – Описательные статистики»*. Для этого следует заполнить

диалоговое окно, приведенное на рис.11 и получить результат в интервале ячеек H7:I23 (рис.14).

	A	B	C	D	E	F	G	H	I	J	K	L	M	N
3	Значения			признака					"Пакет Анализа –		Ранг и перцентиль"			
4	X										Точка	Столбец1	Ранг	Процент
5	-2,18										60	12,04	1	100,00%
6	-0,57							2-й способ			59	9,9	2	98,30%
7	-0,36			1-й способ -				«Пакет Анализа –			58	9,55	3	96,60%
8	-0,3			встроенные ф Excel				Описательные статистики»			57	9,28	4	94,90%
9	0,43								X		56	9,09	5	93,20%
10	0,85										55	9,06	6	91,50%
11	1,33			Среднее		4,64583		Среднее	4,645833333		54	8,44	7	89,80%
12	1,42							Стандартна	0,375927801		53	8,23	8	88,10%
13	1,52			Медиана		4,28		Медиана	4,28		52	8,17	9	86,40%
14	1,58			Мода		4,05		Мода	4,05		51	8,04	10	84,70%
15	2,06			Станд.откл.		2,91192		Стандартно	2,911924224		50	7,48	11	83,00%
16	2,09			Дисперсия		8,4793		Дисперсия	8,479302684		49	7,38	12	81,30%
17	2,27			Эксцесс		-0,0963		Эксцесс	-0,096252528		48	7,21	13	79,60%
18	2,36			Асимметричность		0,11972		Асимметри	0,119723303		47	7,09	14	77,90%
19	2,38			Интервал				Интервал	14,22		46	6,75	15	76,20%
20	2,66			Минимум		-2,18		Минимум	-2,18		45	6,23	16	74,50%
21	2,98			Максимум		12,04		Максимум	12,04		44	6,22	17	72,80%
22	3,54			Сумма				Сумма	278,75		43	6,2	18	71,10%
23	3,65			Объем выборки		60		Счет	60		42	5,92	19	69,40%
49	6,23				Квартили			Персентили			16	2,66	45	25,40%
50	6,75				0	-2,18	мин	0%	-2,18	мин	15	2,38	46	23,70%
51	7,09				1	2,59		5%	-0,303		14	2,36	47	22,00%
52	7,21				2	4,28		50%	4,28		13	2,27	48	20,30%
53	7,38				3	6,36		95%	9,2935		12	2,09	49	18,60%
54	7,48				4	12,04	макс	100%	12,04	макс	11	2,06	50	16,90%
55	8,04										10	1,58	51	15,20%
56	8,17										9	1,52	52	13,50%
57	8,23										8	1,42	53	11,80%
58	8,44										7	1,33	54	10,10%
59	9,06										6	0,85	55	8,40%
60	9,09										5	0,43	56	6,70%
61	9,28										4	-0,3	57	5,00%
62	9,55										3	-0,36	58	3,30%
63	9,9										2	-0,57	59	1,60%
64	12,04										1	-2,18	60	,00%

Рис. 13. Рабочий лист MS Excel в режиме отображения данных.

При выполнении п.5 следует учесть, что функция распределения в данном случае должна быть разрывной. Поскольку наблюдений много (*n*=60), то ступени мелкие (1/*n*=1/60), поэтому будем считать функцию распределения непрерывной (при том, что MS Excel не умеет строить разрывные функции).

	D	E	F
6	**1-й способ -**		
7	**встроенные ф. I**		
8	Среднее		**=СРЗНАЧ(Набор_данных)**
10	Медиана		**=МЕДИАНА(Набор_данных)**
11	Мода		**=МОДА(Набор_данных)**
12	Станд.откл.		**=СТАНДОТКЛОН(Набор_данных)**
13	Дисперсия		**=ДИСП(Набор_данных)**
14	Эксцесс		**=ЭКСЦЕСС(Набор_данных)**
15	Асимметричность		**=СКОС(Набор_данных)**
16	Интервал		
17	Минимум		**=МИН(Набор_данных)**
18	Максимум		**=МАКС(Набор_данных)**
19	Сумма		
20	Объем выборки		**=СЧЁТ(Набор_данных)**

	E	F	G	H	I
49	**Кв**			**Пер**	
50	0	=КВАРТИЛЬ(Набор_данных;E50)		0	=ПЕРСЕНТИЛЬ(Набор_данных;H50)
51	1	=КВАРТИЛЬ(Набор_данных;E51)		0,05	=ПЕРСЕНТИЛЬ(Набор_данных;H51)
52	2	=КВАРТИЛЬ(Набор_данных;E52)		0,5	=ПЕРСЕНТИЛЬ(Набор_данных;H52)
53	3	=КВАРТИЛЬ(Набор_данных;E53)		0,95	=ПЕРСЕНТИЛЬ(Набор_данных;H53)
54	4	=КВАРТИЛЬ(Набор_данных;E54)		1	=ПЕРСЕНТИЛЬ(Набор_данных;H54)

Рис. 14. Фрагмент рабочего листа MS Excel в режиме отображения формул

Построение графика функции распределения проведем в два этапа:

- с помощью надстройки MS Excel *«Пакет Анализа - Ранг и персентиль»* получим таблицу с накопленными частотами. Результат работы надстройки приведен на рис.15 (интервал ячеек K4:N64);

	K	L	M	N
2	**"Пакет Анализа –**			
3	**Ранг и перцентиль"**			
4	*Точка*	*Столбец1*	*Ранг*	*Процент*
5	60	12,04	1	100,00%
6	59	9,9	2	98,30%
63	2	-0,57	59	1,60%
64	1	-2,18	60	,00%

Рис. 15. Фрагмент рабочего листа MS Excel с результатами работы надстройки «Пакет Анализа - Ранг и персентиль»

- полученные значения отсортировать по возрастанию значений вариант в другом диапазоне рабочего листа – в интервале ячеек B132:C194 (рис.16);

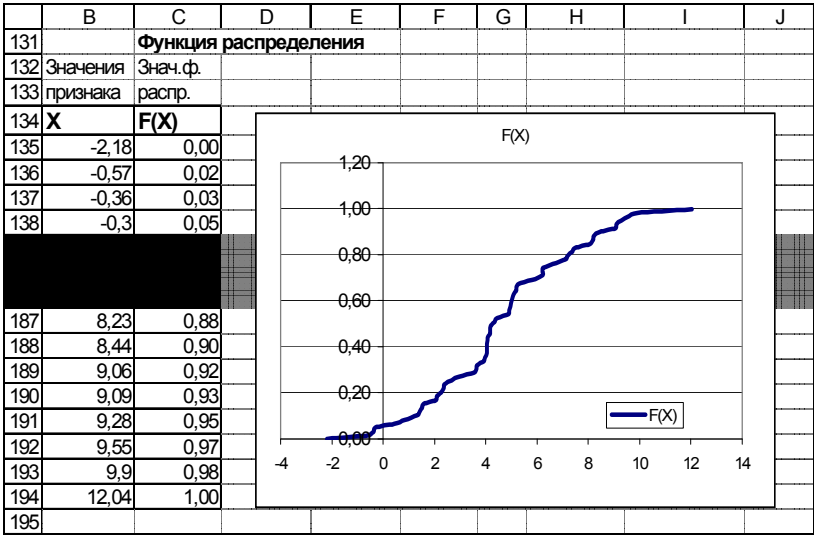


Рис. 16. Построение функции распределения

Назначить числовой формат данных в ячейках C135:C194, построить график эмпирической функции распределения. Результат представлен на рис.16.

Для выполнения п.6 строим диаграмму «Ящик с усами» самостоятельно любым удобным способом (карандаш + линейка + бумага или MS Excel или средства типа AutoCAD и т.д).

Перед построением заполним таблицу (рис.17), предварительно выполнив соответствующие расчеты:

Для заполнения таблицы следует использовать результаты, полученные в п.4, остальные значения (интерквартильный диапазон; внешнее и внутреннее верхнее и нижнее ограждения) вычисляем самостоятельно.

Характеристики "ящика с усами"	
Параметр	Значение
Внешнее верхнее ограждение	
Внутреннее верхнее ограждение	
Верхний ус	
Максимальное значение	
Третья квартиль	
Медиана	
Первая квартиль	
Минимальное значение	
Нижний ус	
Внутреннее нижнее ограждение	
Внешнее нижнее ограждение	
Интерквартильный диапазо	

Рис. 17. Таблица для построения диаграммы «Ящик с усами»

Для выполнения п.7 по формулам (10 – 12) найдем размах, количество и длину интервалов. Результаты вычислений приведены на рис.18. Сам интервальный ряд приведен в ячейках C67:G78 на рис.19. При построении интервального ряда ширина интервала округлена до 2.0, в качестве левой границы первого интервала взято значение –4, что привело к увеличению числа интервалов до 9.

	A	B
67		
68	Размах	
69	14,22	
70		
71	интервалы	
72	кол-во	размер
73	7	2,058814
74		

Рис. 18. Фрагмент рабочего листа MS Excel. Определение параметров интервального вариационного ряда

При выполнении п.8 гистограмму построим двумя способами: с помощью «**Мастера Диаграмм**» и с помощью надстройки MS Excel «**Пакет Анализа→Гистограмма»** (рис.19 - рис.20).

	C	D	E	F	G	H	I	J	K	L
66	**Построение интервального ряда**									
67		Границы инетрвала		середина интервала	частота	Накопленная частота	Относительная частота	Накопленная отн.частота	Среднее	Дисперсия
68		левая	правая							
69	№	a[i]	b[i]	x[i]	mi		mi/N		mi*xi	mi*(xi-x ср)^2
70	1	-4	-2	-3,00	1	1	0,0167	0,0167	-3	60,32
71	2	-2	0	-1,00	3	4	0,0500	0,0667	-3	99,76
72	3	0	2	1,00	6	10	0,1000	0,1667	6	85,13
73	4	2	4	3,00	12	22	0,2000	0,3667	36	37,45
74	5	4	6	5,00	20	42	0,3333	0,7000	100	1,09
75	6	6	8	7,00	8	50	0,1333	0,8333	56	39,90
76	7	8	10	9,00	9	59	0,1500	0,9833	81	161,29
77	8	10	12	11,00	0	59	0,0000	0,9833	0	0,00
78	9	12	14	13,00	1	60	0,0167	1,0000	13	67,79
79	sum				60				286	552,73
80	average								4,77	9,21
81	Станд.откл.									3,035164283

Рис. 19. Фрагмент рабочего листа MS Excel.
Построение интервального вариационного ряда

Обратим внимание, что это средство позволяет получить значения частот, не обращаясь к их непосредственному подсчету.

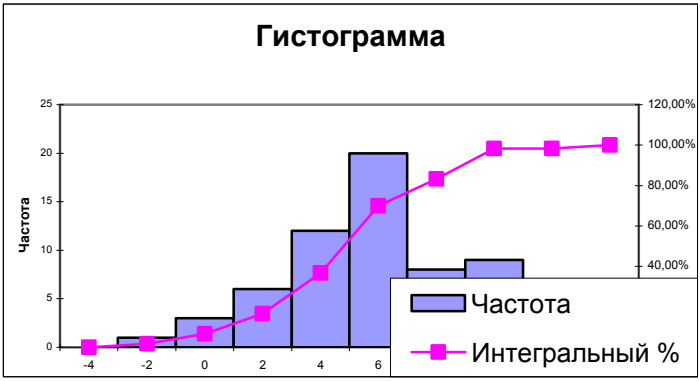


Рис. 20. Диаграмма MS Excel, полученная с помощью надстройки «Пакет Анализа - Гистограмма»

При выполнении п.9 для вычисления значений m_X, D_X, σ_X заполним интервал ячеек K70:L78. Результаты вычислений представлены в ячейках K80, L80-L81.

При выполнении п.10 доверительный интервал вычислим по формулам (22)-(23).

Результаты вычислений приведены на рис.21-22.

	A	B	C	D	E	F	G	H	I	J
197	**Построение доверительного интервала**									
198	**для оценки параметра m (среднего генеральной совокупности),**									
199	**в предположении , что данное распределение нормально**									
200	**с неизвестным стандартным отклонением**									
201	Gamma=	0,95								
202	Вычисление t			2,000997	<—	=t (1-GAMMA; N-1)=СТЬЮДРАСПОБР(0,05;59)				
203	S-исправленное среднеквадратичное отклонение случайной величины X									
204	S=	2,936498								
205	Вычисление DELTA			0,758579	<—	=t*S/(n^(1/2))				
206	интервал									
207		левая граница		правая граница						
208		$\bar{x}_в - \frac{t_{1-\gamma,n-1}\cdot S}{\sqrt{n}} < m_x < \bar{x}_в + \frac{t_{1-\gamma,n-1}\cdot S}{\sqrt{n}}$								
209										
210										
211		3,887255		5,404412						
212										

Рис. 21. Построение доверительного интервала
(режим отображения данных)

	A	B	C	D
199	Gamma=	0,95		
200	Вычислен			=СТЬЮДРАСПОБР(1-B199;Объем-1)
201	S-исправ			
202	S=	=КОРЕНЬ(F13*Объем/(Объем-1))		
203	Вычислен			=D200*B202/КОРЕНЬ(Объем)
204	интервал			
205		левая граница		правая граница
209		=F8-D203		=F8+D203

Рис. 22. Построение доверительного интервала
(режим отображения формул)

СПИСОК ЛИТЕРАТУРЫ

1. *Гмурман В.Е.* Теория вероятностей и математическая статистика, изд.9. - М.: Высшая школа, 2003, с.480.
2. *Господариков А.П.* Математический практикум. Ч.5. Теория вероятности и математическая статистика. Теория функций комплексного переменного. Операционное исчисление. Теория поля. - СПб: СПГГИ(ТУ), 2003, с.187
3. *Бер К., Кэйри П.* Анализ данных с помощью Microsoft Excel. - М.: Вильямс, 2004, с. 560.

ОГЛАВЛЕНИЕ

ЭКОНОМИКО-МАТЕМАТИЧЕСКИЕ МЕТОДЫ И МОДЕЛИРОВАНИЕ

ОПИСАТЕЛЬНАЯ СТАТИСТИКА

Методические указания к лабораторным работам для студентов бакалавриата направления 21.03.02

Сост.: *В.В. Беляев, Т.Р. Косовцева*

Печатается с оригинал-макета, подготовленного кафедрой информатики и компьютерных технологий

Ответственный за выпуск *В.В. Беляев*

Лицензия ИД № 06517 от 09.01.2002

Подписано к печати 03.11.2020. Формат 60×84/16.
Усл. печ. л. 2,0. Усл.кр.-отт. 2,0. Уч.-изд.л. 1,8. Тираж 75 экз. Заказ 821.

Санкт-Петербургский горный университет
РИЦ Санкт-Петербургского горного университета
Адрес университета и РИЦ: 199106 Санкт-Петербург, 21-я линия, 2

Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное учреждение высшего образования
Санкт-Петербургский горный университет

Кафедра информатики и компьютерных технологий

ЭКОНОМИКО-МАТЕМАТИЧЕСКИЕ МЕТОДЫ И МОДЕЛИРОВАНИЕ

ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Методические указания к лабораторным работам для студентов бакалавриата направления 21.03.02

САНКТ-ПЕТЕРБУРГ
2020

УДК 519.25 (073)

ЭКОНОМИКО-МАТЕМАТИЧЕСКИЕ МЕТОДЫ И МОДЕЛИРОВАНИЕ. Проверка статистических гипотез: Методические указания к лабораторным работам / Санкт-Петербургский горный университет. Сост.: *В.В. Беляев, Т.Р. Косовцева*. СПб, 2020. 39 с.

Приведены необходимые теоретические сведения и примеры выполнения заданий по некоторым разделам теории вероятностей и математической статистики, которые являются теоретической основой применения корреляционно-регрессионного анализа. Все решения выполнены с использованием электронных таблиц MS Excel, в том числе с применением надстройки «Пакет анализа».

Предназначены для студентов бакалавриата направления 21.03.02 «Городской кадастр», а также быть могут использованы и для студентов других направлений, изучающих статистику.

Научный редактор доц. *А.Б. Маховиков*

Рецензент канд. техн. наук *К.В. Столяров* (Телум Инк)

ВВЕДЕНИЕ

Корреляционно-регрессионный анализ является одним из самых применяемых математических методов при решении задач, возникающих при рассмотрении проблем городского кадастра. Например, такой проблемой является массовая оценка объектов. Одним из краеугольных камней, лежащих в основе корреляционно-регрессионного анализа, является математическая статистика. *Математическая статистика* является частью статистики (от лат. *status* - состояние) - науки, изучающей, обрабатывающей и анализирующей количественные данные о самых разнообразных массовых явлениях окружающей нас жизни.

Одной из основных задач математической статистики является проверка статистических гипотез.

К числу таких гипотез относятся гипотезы относительно законов распределения или значений их параметров.

В подавляющем большинстве реальных ситуаций проверяемая статистическая гипотеза является гипотезой об отсутствии того или иного эффекта:

- об отсутствии различий, например, о равенстве нулю разности средних;
- об отсутствии тех или иных эффектов, связей, соответствий, зависимостей и т.п.

В настоящей работе описано применение этих методов с использованием электронных таблиц MS Excel.

ЛАБОРАТОРНАЯ РАБОТА

ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Цель: Получить понятия о проверке статистических гипотез о виде распределения, о равенстве средних значений и дисперсий двух нормально распределенных совокупностей.

1. БАЗОВЫЕ ПОНЯТИЯ

Часто в процессе исследований требуется по выборочным данным определить закон распределения генеральной совокупности. При этом есть основания предполагать, что этот закон распределения имеет определенный вид. В других случаях закон распределения совокупности известен, и нужно оценить величину параметров. При этом предполагают, что неизвестные параметры распределения равны некоторым определенным значениям.

Пусть H_0 (нулевая гипотеза) – выдвинутая гипотеза о виде распределения или о значении параметров распределения. Вместе с гипотезой H_0 всегда рассматривается альтернативная (конкурирующая) гипотеза H_1. Гипотеза H_1 часто, но не всегда, является противоположной проверяемой гипотезе H_0.

Статистическим критерием называют случайную величину T, которая используется для проверки нулевой гипотезы.

Множество возможных значений T разбито на два непересекающихся подмножества. Одно из них содержит значения критерия, при которых нулевая гипотеза принимается, другое – при которых H_0 отвергается и принимается альтернативная гипотеза H_1. *Критической областью* называется совокупность значений критерия, при которых гипотезу отвергают. Область принятия гипотезы H_0 - совокупность значений критерия, при которых гипотезу H_0 принимают. Критическая точка $T_{крит.}$ отделяет критическую область от области принятия гипотезы.

Уровень значимости α - достаточно малая вероятность (0.05; 0.01). Значение $T_{крит.}$ вычисляется таким образом, чтобы при справедливости гипотезы H_0 вероятность того, что значение T превзойдет значение $T_{крит}$ была равна α. *Таким образом при проверке многих статистических гипотез (но не всех), следуют правилу: если*

найденное значение T меньше $T_{крит}$ ($T < T_{крит}$), то гипотезу H_0 принимают, в противоположном случае отвергают и принимают гипотезу H_1 .

При проверке статистических гипотез возможны четыре исхода, в том числе могут быть допущены ошибки двух родов (табл.1):

Таблица 1

Возможные результаты проверки статистической гипотезы H_0

Гипотеза H_0	H_0 *- верная гипотеза*	H_0 *-ложная гипотеза*
H_0 -отвергается	Верная гипотеза H_0 отвергается (ошибка первого рода), событие *A*.	Ложная гипотеза H_0 отвергается (правильное решение)
H_0 -принимается	Верная гипотеза H_0 принимается (правильное решение)	Ложная гипотеза H_0 принимается (ошибка второго рода), событие *B*.

Ошибка первого рода возникает, когда отвергается гипотеза H_0, хотя она верна. Уровень значимости α - это вероятность того, что допускается ошибка первого рода, т.е. $P(A) = \alpha$. Чем меньше уровень значимости, тем меньше вероятность отвергнуть верную гипотезу. Обычно значение α полагают равным 0.05 (5%-ный уровень значимости) или 0.01 (1%-ный уровень значимости).

Ошибка второго рода возникает, когда принимается гипотеза H_0, когда она неверна. Вероятность недопущения ошибки второго рода называется *мощностью критерия*, эта вероятность равна $1\text{-}P(B)$. Мощность критерия – вероятность того, что гипотеза H_0, будет принята, когда верна альтернативная H_1.

Результат проверки не доказывает истинность или ложность гипотезы H_0, но говорит лишь о том, что экспериментальные данные противоречат или не противоречат этой гипотезе.

На практике часто производится проверка следующих гипотез:

- о предполагаемом законе распределения неизвестного распределения генеральной совокупности, из которой получена выборка (критерий согласия Пирсона χ^2 (хи–квадрат));
- о равенстве дисперсий двух нормально распределенных генеральных совокупностей (критерий Фишера);
- о равенстве двух средних нормально распределенных генеральных совокупностей (критерий Стьюдента (*t*–критерий).

Критерием согласия называют критерий для проверки гипотезы о предполагаемом законе неизвестного распределения. Наиболее распространен критерий согласия Пирсона χ^2 («хи–квадрат»). При использовании критерия «хи–квадрат» предполагают, что выборочные данные получены из генеральной совокупности с известным законом распределения (гипотеза H_0). Альтернативной гипотезой является H_1: выборочные данные получены из генеральной совокупности с другим законом распределения.

При проверке по этому критерию находят теоретические частоты, сравнивают их с эмпирическими частотами. Если расхождение случайно, то гипотезу H_0 принимают, иначе принимают гипотезу H_1 .

Пусть N наблюдений распределены по k разрядам, m_i – количество наблюдений в i-том разряде. Пусть известен закон распределения генеральной совокупности, из которой предположительно получена выборка, и p_i - вероятность попадания в i-ый разряд. Тогда $N \cdot p_i$ - теоретическая частота i-го разряда.

Меру расхождения между эмпирическими (фактическими) частотами и предполагаемыми теоретическими определяют как:

$$\chi^2 = \sum_i^k \frac{(m_i - N \cdot p_i)^2}{N \cdot p_i}, \qquad (1)$$

где k – количество разрядов, в которые сведены результаты опытов; m_i – количество наблюдений в i-том разряде; p_i – теоретическая вероятность (в соответствии с предполагаемым законом распределения) i-го разряда.

Далее определяется число степеней свободы r:

$$r = k - c - 1, \qquad (2)$$

где: c – число параметров теоретического распределения. Для закона распределения Пуассона и показательного закона распределения $c = 1$; для нормального, непрерывного равномерного – $c = 2$; для непрерывного дискретного – $c = 0$.

По заданным значениям α и r с помощью специальной таблицы находят $\chi^2_{крит.}$ В случае отсутствия таблиц можно

использовать встроенную функцию MS Excel ХИ2ОБР, которая имеет следующий формат:

ХИ2ОБР (вероятность; число степеней свободы)

например: ХИ2ОБР(0.05; 5).

Сравнивают найденное значение χ^2 с значением $\chi^2_{крит}$.

Если $\chi^2 \le \chi^2_{крит,}$ то гипотезу H_0 о совпадении эмпирического распределения с теоретическим принимают.

В противном случае ($\chi^2 > \chi^2_{крит}$), гипотезу H_0 отвергают, принимают гипотезу H_1.

Практически m_i должны быть больше или равны 5, если в некотором интервале это условие нарушается, то интервал объединяется с соседним.

Пример 1

В результате 300 бросков кости (кубика с 6 гранями) были получены следующие результаты (табл.2):

Таблица 2

Число очков	1	2	3	4	5	6
Кол-во повторений	38	38	50	52	57	65

Проверить гипотезу о том, что кость «правильная».

Решение

Если кость «правильная», то в таблице количество повторений числа выпавших очков должно быть почти одинаково. В этом случае имеет место равномерный закон распределения числа выпавших очков, и каждый из шести возможных исходов (число выпавших очков) равновероятен, т.е. P_i=1/6 для всех i=1,2…6.

Таким образом, задача оценки «правильности» кости сводиться к проверке гипотезы H_0, состоящей в том, что выборка, приведенная в табл.2, отобрана из равномерно распределенной генеральной совокупности. Альтернативной является гипотеза H_1 – данная выборка не является равномерно распределенной, точнее отобрана из совокупности распределенной, имеющей не равномерное распределение. Все расчеты сведены в табл.3.

Получено значение χ^2 =11,320. В нашем случае r = 6-1= 5, и при α=0.05 находим значение $\chi^2_{крит}$ = 11,07. Поэтому гипотеза H_0 о равномерном распределении отвергается и принимается альтернативная гипотеза H_1.

Таблица 3

Xi	***mi***	***P_i***	***N*Pi***	***(mi-N*Pi)^2/N*Pi***	
1	38	0,1667	50	2,88	
2	38	0,1667	50	2,88	
3	50	0,1667	50	0,00	
4	52	0,1667	50	0,08	
5	57	0,1667	50	0,98	
6	65	0,1667	50	4,50	
N=	300			11,32	*Xi-квадрат*

Приходим к выводу, что кость «неправильная». На рис.1. представлено сравнение экспериментального и теоретического распределений числа выпавших очков при бросании игральной кости.

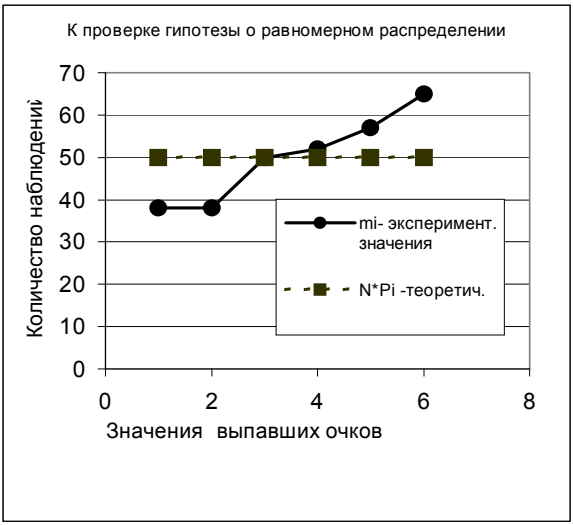


Рис. 1. Сравнение экспериментального и теоретического распределений числа выпавших очков при бросании игральной кости.

Пример 2

Проверить гипотезу о том, что распределение, представленное в таблице 4, является частным случаем нормального распределения (уровень значимости $\alpha = 0,05$)

Таблица 4

Интервал	Частота mi
[-4;-2]	1
[-2;0]	3
[0;2]	6
[2;4]	12
[4;6]	20
[6;8]	8
[8;10]	8
[10;12]	1
[12;14]	1

Решение

Построим полигон распределения (рис.2). Кривая имеет колоколообразную форму и симметрична, поэтому можно сделать предположение, что выборка получена из нормально распределенной генеральной совокупности. Выдвигаем гипотезу H_o: данное эмпирическое распределение следует нормальному закону распределения. Альтернативная гипотеза – H_1: эмпирическое распределение не следует нормальному закону распределения.

Воспользуемся результатами, полученными [4, рис.19].

Будем считать, что диапазон ячеек C67:L81 уже заполнен, т.е. уже вычислены среднее значение x_{cp}, равное 4,77, и стандартное отклонение σ, равное 3,03:

- в интервале ячеек D70:E78 указаны значения левой и правой границ интервала;
- в интервале ячеек F70:F78 – середины интервалов;
- в интервале ячеек G70:G78 - наблюдаемые значения частот m_i .

Случайная величина называется распределенной по нормальному закону, если она имеет следующую плотность распределения (3).

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-m)^2}{2\sigma^2}}, \qquad (3)$$

где m и σ – параметры распределения. Причем m - математическое ожидание, σ^2 -дисперсия случайной величины.

Оценкой параметра m генеральной совокупности является выборочное среднее $\bar{x}_в$, в данном случае $\bar{x}_в = 4{,}77$.

Оценкой параметра σ является среднее квадратическое отклонение $\sigma_в$, равное *3,04* .

Дальнейшие вычисления показаны на рис.2.*а* и рис.2.*б*.

Вычислим p_i - вероятность того, что случайная величина, имеющая нормальное распределение с параметрами $m = 4{,}77$ и $\sigma = 3.04$ попадает в *i*-ый интервал.

Для этого воспользуемся соотношением:

$$p_i = P(a_i \le x \le b_i) = F(b_i) - F(a_i), \qquad (4)$$

где $F(x)$ – функция распределения. В MS Excel для нормального закона распределения с параметрами m и σ значение $F(x)$ можно найти, воспользовавшись встроенной функцией НОРМРАСП:

$F(x)$ = НОРМРАСП(x; m; σ; ИСТИНА).

Выполнение:

- В ячейки N70:N78 заносим значения $F(b_i)$;
- В ячейки M70:M78 заносим значения $F(a_i)$;
- В интервале O70:O78 получаем теоретические значения вероятности, вычисленные по формуле (4);

	C	D	E	F	G	K	L	M	N
67		Границь	интервал	середина		Среднее	Дисперсия		
68		левая	правая	интерва	частота	относ.ч.			
69	№	a[i]	b[i]	x[i]	mi	mi*xi	mi*(xi-x ср	F(a[i])	F(b[i])
70	1	-4	-2	-3.00	1	-3	60.32	0.001936	0.012893
71	2	-2	0	-1.00	3	-3	99.76	0.012893	0.058152
72	3	0	2	1.00	6	6	85.13	0.058152	0.181006
73	4	2	4	3.00	12	36	37.45	0.181006	0.400291
74	5	4	6	5.00	20	100	1.09	0.400291	0.657757
75	6	6	8	7.00	8	56	39.90	0.657757	0.856628
76	7	8	10	9.00	9	81	161.29	0.856628	0.957667
77	8	10	12	11.00	0	0	0.00	0.957667	0.991418
78	9	12	14	13.00	1	13	67.79	0.991418	0.998825
79	sum				60	286	552.73		
80	average					4.77	9.21		
81	Станд.откл.						3.035164		

Рис. 2.*а*. Проверка гипотезы о нормальности эмпирического распределения (начало).

	O	P	Q	R	S	T
67	Вероятность	Теоретич	Объединенны		Хи-квадрат	
68	теорегич.	частоты	интервалы			
69	pi=F(b[i])-F(a[i])	=N*pi	ni	=N*pi	Chi_squ	
70	0.0109565	0.657391				
71	0.0452589	2.715531				
72	0.1228544	7.371262	10	10.744	0.0515	
73	0.2192847	13.15708	12	13.157	0.1018	
74	0.2574658	15.44795	20	15.448	1.3414	
75	0.1988713	11.93228	8	11.932	1.2959	
76	0.1010394	6.062363	10	8.5318	0.2526	
77	0.0337508	2.025049				
78	0.0074073	0.444438				
79	0.99688903	59.81334	60	59.813	3.0432	наблюда
80						
81					5.9915	критичес

Рис. 2.*б*. Проверка гипотезы о нормальности эмпирического распределения (окончание).

- В интервал P70:P78 заносим теоретические значения частот, вычисленные по формуле $N \cdot p_i$.

Для визуального сравнения наблюдаемых m_i и теоретических $N \cdot p_i$ частот построим соответствующие графики – полигоны частот, по оси абсцисс отложим середины интервалов (рис.3). Заметим, что графики схожи: симметричны относительно оси,

проходящей через максимум, и плавно убывают по мере удаления от максимума.

Как следует из рис.2.*б*, значения теоретических частот в ряде интервалов меньше 5. Такие интервалы обычно объединяют. Объединим интервалы 1 - 3 и 7- 9 включительно, суммируя значения наблюдаемых и теоретических частот. Найденные суммы занесем в ячейки Q72:R72 и Q76:R76.

Объединим интервалы с 7 по 9 включительно, соответствующие значения наблюдаемых и теоретических частот также суммируются, результат в ячейках Q76:R76.

Ячейки Q73:R75 заполняют обычным образом.

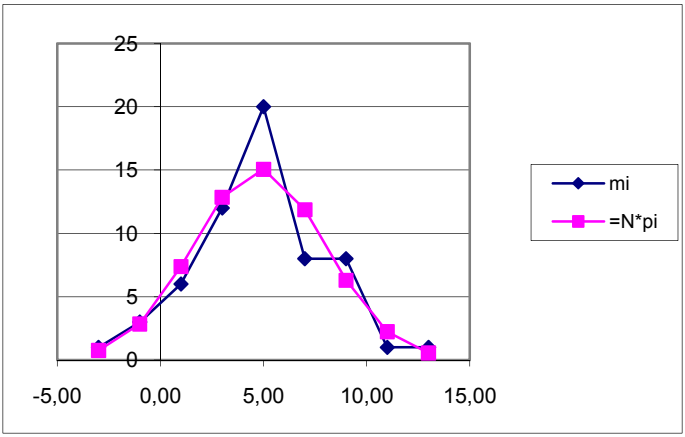


Рис. 3. Сравнение экспериментального и соответствующего нормального (теоретического) распределений

После объединения ячеек число интервалов сократилось до 5, т.е. теперь k=5.

Вычислим значение χ^2 по формуле (1). Для этого в ячейку S72 заносим формулу

=(Q72-R72)^2/R72,

копируем её в ячейки диапазона S72:S76.

Сумму чисел из ячеек S72:S76 помещаем в ячейку S79 – это искомое значение «хи–квадрат» ($\chi^2 = 3.04$.).

Определим число степеней свободы r по формуле (2). Поскольку k=5 , c=2, то r = 2.

Для $\alpha = 0{,}05$ и $r = 2$ найдем $\chi^2_{крит}$, используя встроенную функцию MS Excel ХИ2ОБР. Тогда $\chi^2_{крит}$ =ХИ2ОБР(0.05;2) = *5.99*

Поскольку найденное значение $\chi^2 < \chi^2_{крит}$, то гипотезу H_0 о соответствии распределения случайной величины x нормальному закону распределения принимаем. Согласно выборочным экспериментальным данным нет оснований отвергнуть гипотезу H_0.

2 КРИТЕРИЙ ФИШЕРА

F - критерий Фишера используют для сравнения дисперсий двух генеральных совокупностей, распределенных по нормальному закону.

По независимым выборкам объема из этих совокупностей найдены выборочные дисперсии σ_1^2 и σ_2^2 . Выдвигается гипотеза H_0 - дисперсии равны, альтернативная гипотеза H_1 - дисперсии не равны.

Вычисляется $F_{\text{набл.}}$ по формуле:

$$F_{\text{набл.}} = \frac{\sigma_1^2}{\sigma_2^2}, \qquad (5)$$

где σ_1^2 - большая дисперсия, σ_2^2 - меньшая дисперсия.

По заданному уровню значимости α и числам степеней свободы r_1 и r_2 (r_1 число степеней свободы числителя и r_2 число степеней свободы знаменателя) - определяем $F_{\text{крит.}}$ по таблицам или используя встроенные функции MS Excel.

Число степеней свободы числителя определяется по формуле:

$$df_1 = n_1 - 1, \qquad (6)$$

где n_1- число вариант для большей дисперсии.

Число степеней свободы знаменателя определяется по формуле:

$$df_2 = n_2 - 1, \qquad (7)$$

где n_2 - число вариант для меньшей дисперсии.

Если $F_{\text{набл.}} \le F_{\text{крит.}}$ (вычисленное значение критерия $F_{\text{набл.}}$ ***не*** больше критического), то принимается гипотеза H_0 (дисперсии

равны), в противном случае ($F_{\text{набл.}} > F_{\text{крит.}}$) принимается гипотеза H_1 (дисперсии различны).

Пример 3

При проведении тестирования двух одинаковых приборов были проведены измерения эталона. При этом первым прибором было проведено n_1 =13 измерений, а вторым - n_2 =15.

Результаты были записаны в виде отклонений от значения эталона. Требуется выяснить: одинаковой ли точностью обладают приборы.

Решение

Величина отклонений от эталонного значения для первого прибора (n_1 =13) внесена в столбец ***B*** рабочего листа книги MS Excel**,** а для второго прибора (n_2=15) результаты - в столбец ***C*** (рис.4 - рис.5).

Средние значения отклонений одинаковы и равны нулю. Следовательно, у приборов отсутствует систематическая ошибка.

Проверка точности приборов сводится к проверке совпадения дисперсий. Если дисперсии отклонений от эталонного значения статистически равны, то приборы обладают одинаковой точностью.

Выдвигается гипотеза H_0 - дисперсии выборок равны, альтернативная гипотеза H_1 - дисперсии не равны.

В результате расчета были получены соответственно следующие значения дисперсий: σ_1^2 =60,67 (ячейка B25) и σ_2^2 =20,00 (ячейка C25).

Значение критерия $F_{\text{набл.}}$ вычислим в ячейке B27 по формуле =60,67 /20,00 = 3,0333.

Для уровня значимости α =0.05; числа степеней свободы числителя df_1 =13-1=12 и числа степеней свободы знаменателя df_2 = 15-1= 14 находим $F_{\text{крит}}$ с помощью встроенной функции FРАСПОБР(). $F_{крит}$= 2,5342.

Поскольку $F_{\text{набл.}} = 3{,}0333 > F_{\text{крит.}} = 2{,}5342$ то гипотеза H_0 *отклоняется*, и принимается альтернативная гипотеза H_1 (дисперсии различны).

Следовательно, приборы имеют различную точность.

	A	B	C	D	E
1	**Критерий Фишера F-критерий)**				
2	**Сравнение двух независимых выборок**				
3	Гипотеза H_0 : **дисперсии выборок равны**				
4	альтернативная гипотеза Н1: дисперсии не равны				
5	№№ измерения	Первый прибор	Второй прибор		
6		x	y		
7	1	-12	-7		
8	2	-10	-6		
9	3	-8	-5		
10	4	-6	-4		
11	5	-4	-3		
12	6	-2	-2		
13	7	0	-1		
14	8	2	0		
15	9	4	1		
16	10	6	2		
17	11	8	3		
18	12	10	4		
19	13	12	5		
20	14		6		
21	15		7		
22	Сумма	0	0		
23	Объем выборки	13	15		
24	Ср.значение	0	0		
25	Дисперсия	60,667	20		
27	Fнабл=	3,03333		alfa=	0,05
29	$F_{\text{набл.}} > F_{\text{крит}}$			Fкрит=	2,53424
31	Гипотеза Н0 отклоняется, принимаем гипотезу Н1.				

Рис. 4. Сравнение двух выборочных дисперсий
(фрагмент рабочего листа MS Excel в режиме отображения данных)

	A	B	C	D
1	**Критерий Фи**			
2	**Сравнение двух**			
3	Гипотеза H_0 : д			
4	альтернативная			
5	№№ измерения	Первый прибор	Второй прибор	
6		x	y	
7	1	-12	-7	
8	2	-10	-6	
20	14		6	
21	15		7	
22	Сумма	=СУММ(B7:B19)	=СУММ(C7:C21)	
23	Объем выборки	=СЧЁТ(B7:B19)	=СЧЁТ(C7:C21)	
24	Ср.значение	=B22/B23	=C22/C23	
25	Дисперсия	=ДИСП(B7:B19)	=ДИСП(C7:C21)	
27	Fнабл=	=B25/C25		
29	alfa=	0,05		
30				
31	Fкрит=	=FРАСПОБР(B29;B23-1;C23-1)		

Рис. 5. Сравнение двух выборочных дисперсий
(фрагмент рабочего листа MS Excel в режиме отображений формул)

Средство анализа «Двухвыборочный F-тест для дисперсии» надстройки «Пакет анализа» MS Excel

Средство анализа «Двухвыборочный F-тест для дисперсии» надстройки «Пакет анализа» MS Excel служит для проверки гипотезы о равенстве дисперсий двух выборок. Для проверки необходимо заполнить диалоговое окно, приведенное на рис.6, назначение всех полей ввода очевидно.

Результаты расчета представлены на рис.7.

Сравним полученные результаты с результатами, полученными вручную.

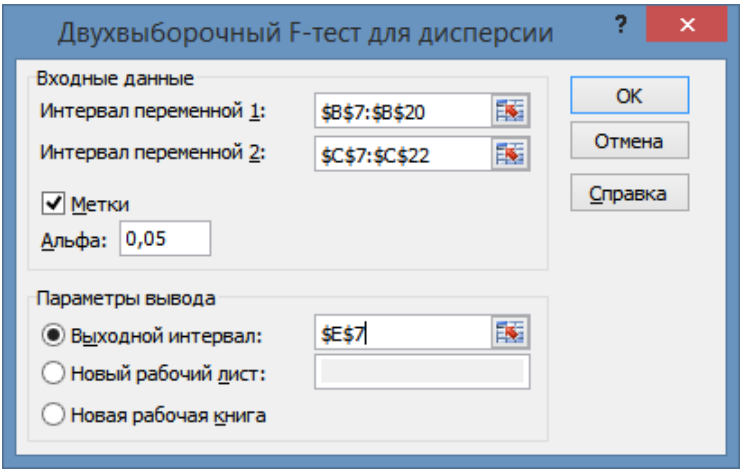


Рис. 6. Диалоговое окно средства анализа «Двухвыборочный F-тест для дисперсии» надстройки «Пакет анализа» MS Excel

	D	E	F	G	H
7		Двухвыборочный F-тест для дисперсии			
8					
9			*x*	*y*	
10		Среднее	0	0	
11		Дисперсия	60,67	20,00	
12		Наблюдения	13	15	
13		df	12	14	
14		F	3,0333		
15		P(F<=f) одностороннее	0,0256		
16		F критическое одностороннее	2,5342		
17					

Рис. 7. Результат применения инструмента «Двухвыборочный F-тест для дисперсии» надстройки «Пакет анализа» MS Excel

При анализе результатов следует помнить, что значение $F_{\text{набл.}}$ (3,0333), которое содержится в ячейке F14, должно быть не меньше единицы. В противоположном случае адреса ячеек с исходными данными следует поменять местами.

Значение $F_{\text{крит.}}$ (равное 2,5342) содержится в ячейке F16. Поскольку содержимое ячейки F14 больше содержимого ячейки F16, то приходим к выводу, совпадающему с выводом, полученным ранее: гипотеза H_0 *отклоняется*, и принимается альтернативная гипотеза H_1 (дисперсии различны).

Тот же вывод (даже более глубокий) можно получить, анализируя содержимое ячейки F15. Конкретно в этом тесте это значение названо «P(*F*<=*f*) одностороннее», но в других тестах подобную величину называют «*p* – значение» или «*p* - value».

Величина p – значение имеет следующий вероятностный смысл: p – значение - это вероятность допустить ошибку первого рода в случае когда верна гипотеза H_0, а наблюдаемое значение критерия принимает значение больше $F_{\text{набл.}}$. Т.е. это вероятность события $P(X > F_{\text{набл.}})$, где случайная величина X имеет *F*-распределение, т.е. $X \sim F(df_1, df_2)$.

Приведенное выше утверждение имеет простую геометрическую трактовку: $p-$ значение - это площадь правого хвоста функции плотности случайной величины $X \sim F(df_1, df_2)$, хвост начинается с $F_{\text{набл.}}$ (рис.8).

Вспомним геометрическую трактовку связи между $F_{\text{крит.}}$ и уровнем значимости α : площадь правого хвоста функции плотности случайной величины $X \sim F(df_1, df_2)$ равна α, хвост начинается с $F_{\text{крит.}}$ (рис. 9).

Проведем сравнение на рис.10 площади хвостов подграфиков, совместив показанные на рисунках 8 и 9. Очевидно, что будет справедливо следующее утверждение.

Неравенство $F_{\text{набл.}} > F_{\text{крит.}}$ равносильно неравенству «*p* – значение меньше уровня значимости α ».

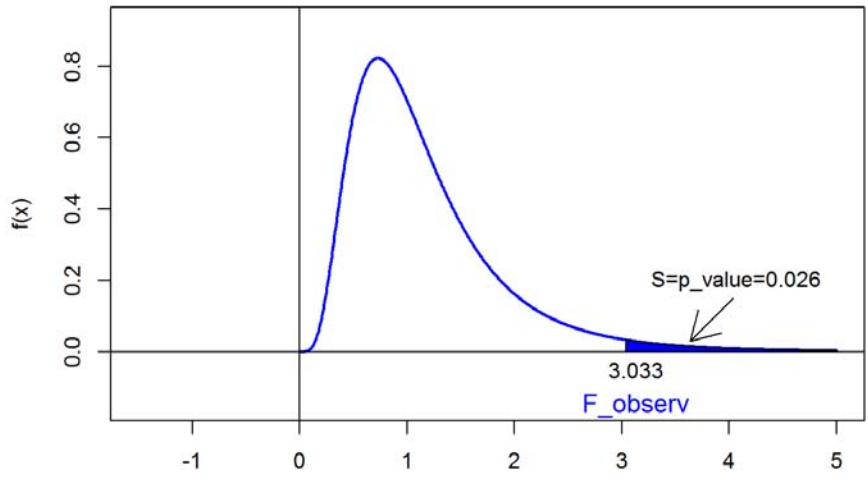


Рис. 8. Геометрический смысл величины *p* – значение и ее связь с величиной $F_{\text{набл.}}$

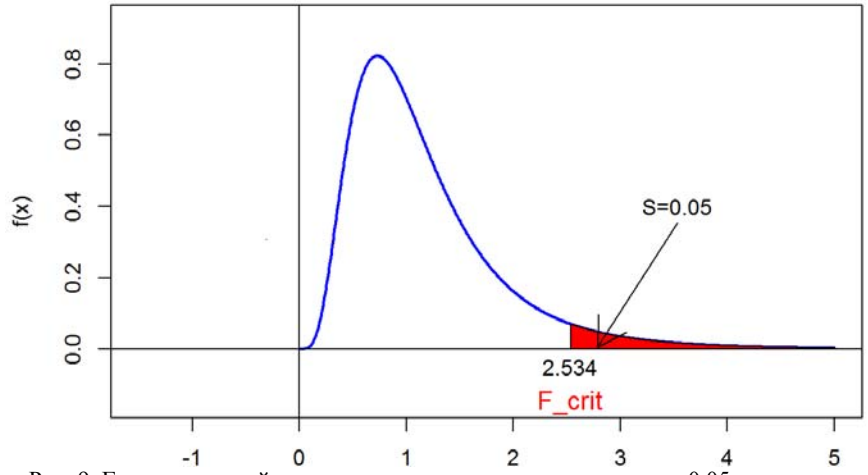


Рис. 9. Геометрический смысл величины уровня значимости α =0.05 и ее связь с величиной $F_{\text{крит.}}$

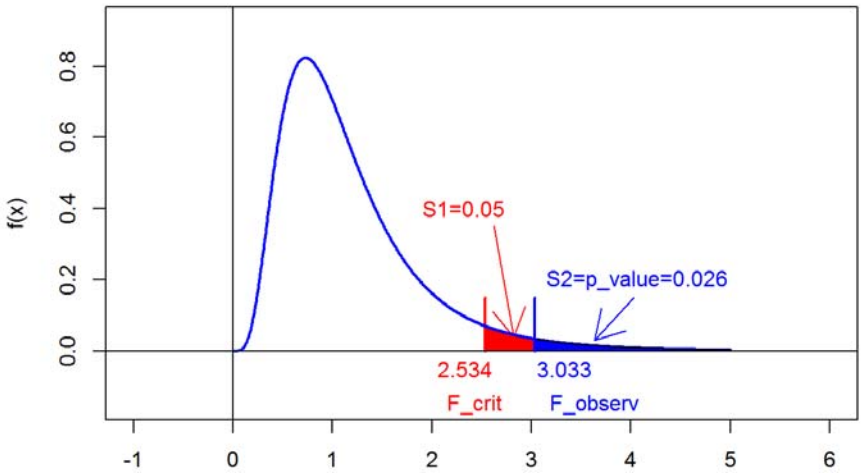


Рис. 10. Сравнение площадей хвостов подграфика.

Аналогично, справедливо утверждение: неравенство $F_{\text{набл.}} < F_{\text{крит.}}$ равносильно неравенству «p – значение больше уровня значимости α ».

Приведенное выше рассуждение позволяет прийти к выводу гипотеза H_0 принимается если «p – значение больше уровня значимости α »; альтернативная гипотеза H_1 принимается если «p – значение меньше уровня значимости α ».

Указанный механизм реализован в большинстве математических пакетов, в надстройке «Пакет анализа» MS Excel в том числе. Большим плюсом этого механизма является отсутствие необходимости вычисления критического значения и возможность оценки величины уровня значимости, при котором гипотеза будет подтверждена или отвергнута.

3 КРИТЕРИЙ СТЬЮДЕНТА (*t*- КРИТЕРИЙ)

Критерий используется для проверки гипотезы о равенстве средних значений двух выборок, взятых из нормально распределенных совокупностей.

Пусть заданы две генеральные совокупности ***x*** и ***y***, имеющие нормальное распределение, из них взяты выборки $\{x_i\}_{i=1}^{n_1}$ и $\{y_i\}_{i=1}^{n_2}$, т.е. n_1 и n_2 - объемы первой и второй выборки соответственно. Выдвигается гипотеза H_0, что средние значения выборок равны (альтернативная гипотеза H_1 - средние значения не равны).

Значение $t_{набл}$ вычисляют по формуле:

$$t_{набл} = \frac{\bar{x} - \bar{y}}{S}, \qquad (8)$$

где $\bar{x}$, $\bar{y}$ — средние арифметические выборок $\{x_i\}_{i=1}^{n_1}$ и $\{y_i\}_{i=1}^{n_2}$;

S - стандартная ошибка разности средних значений.

Число степеней свободы вычисляют по формуле:

$$df = n_1 + n_2 - 2. \qquad (9)$$

Из таблиц для заданного уровня значимости α и числа степеней свободы k определяют $t_{крит}$ (критическое значение).

Если $|t_{набл}| < t_{крит}$, то гипотеза H_0 принимается, в противном случае принимается альтернативная гипотеза.

Стандартная ошибка разности средних значений S вычисляется различными способами в зависимости от поставленной задачи:

- сравнение двух выборок;
- сравнение двух зависимых выборок;
- сравнение более двух независимых выборок.

3.1 СЛУЧАЙ ДВУХ НЕЗАВИСИМЫХ ВЫБОРОК

Требуется сравнить средние значения двух независимых выборок. Здесь возможны два варианта:

1. Дисперсии выборок равны.
2. Дисперсии выборок не равны.

Рассмотрим первый вариант (дисперсии выборок равны). В этом случае значение S вычисляется по формуле:

$$S = \sqrt{\frac{\sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}, \qquad (10)$$

где n_1 и n_2 - объемы первой и второй выборки; $\bar{x}$ и $\bar{y}$ — средние арифметические выборок.

Пример 4

В двух группах учащихся — экспериментальной и контрольной — применялись две различные методики обучения: экспериментальная и традиционная. После завершения обучения был проведен тест и получены следующие результаты по учебному предмету (тестовые баллы; см. табл.4).

Таблица 4

Результаты эксперимента

Первая группа (экспериментальная), N_1=11 человек										
12	14	13	16	11	9	13	15	15	18	14
Вторая группа (контрольная), N_2=9 человек										
13	9	11	10	7	6	8	10	11		

Имеет ли экспериментальный метод обучения преимущество по сравнению с традиционным?

Решение

Для выявления преимущества экспериментального метода обучения по сравнению с традиционным можно проверить совпадения средних оценок в двух группах. Если средние отличаются незначимо, то преимущества нет, в противном случае преимущество есть.

Гипотеза H_0: средние значения выборок равны, альтернативная гипотеза H_1 : средние значения не равны.

Решение приведено на рис.11 - 12.

Общее количество членов выборки: n_1=11, n_2=9; средние значения: $\bar{\boldsymbol{x}}$=13,636; $\bar{\boldsymbol{y}}$=9,444.

По формуле (10) находим стандартную ошибку разности средних значений:

$$S = \sqrt{\frac{\sum (x_i - \bar{x})^2 + \sum (y_i - \bar{y})^2}{n_1 + n_2 - 2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = \sqrt{\frac{60.545 + 38.222}{11 + 9 - 2}\left(\frac{1}{11} + \frac{1}{9}\right)} = 1.053$$

	A	B	C	D	E	F	G
1	**Критерий Стьюдента (t-критерий)**						
3	**Сравнение двух независимых выборок**						
4		**x**	**(x-xcp)^2**	**y**	**(y-ycp)^2**		
5		12	2.677686	13	12.641975		
6		14	0.132231	9	0.1975309		
7		13	0.404959	11	2.4197531		
8		16	5.586777	10	0.308642		
9		11	6.950413	7	5.9753086		
10		9	21.49587	6	11.864198		
11		13	0.404959	8	2.0864198		
12		15	1.859504	10	0.308642		
13		15	1.859504	11	2.4197531		
14		18	19.04132				
15		14	0.132231				
16	**Сумма**	150	60.54545	85	38.222222		
17	**Объем выборки**	11		9			
18	**Ср.значение**	13.6363636		9.444444			
19	**Дисперсия**	6.05454545		4.777778			
20							
21	**Вспомогательная проверка, чтобы выснить какой t-тест следует применять.**						
22	**Гипотеза** Hip_0 **:** дисперсии выборок равны						
23	**альтернативная гипотеза Hip1: дисперсии не равны**						
24	**Fнабл=**	1.26723044		**alfa=**	0.05		
25	$F_{набл.} < F_{крит.}$						
26				**Fкрит=**	3.3471679		
27							
28	**Принимаем гипотезу Hip0 => можем использовать t-тест с равными дисперсиями**						
30							
31	**Гипотеза** H_0 **:** средние значения выборок равны						
32	**альтернативная гипотеза H1: средние значения не равны**						
33	**Используем критерий Стьюдента для проверки основной гипотезы H0.**						
34							
35	**S=**	1.05285501		**k=**	18		
36	**tнабл=**	3.98147813		**alfa=**	0.05		
38	$\lvert t_{набл} \rvert > t_{крит}$			**tкрит=**	2.1009237		
40	**Гипотеза H0 отклоняется, принимаем гипотезу H1.**						

Рис. 11. Проверка гипотезы о совпадении двух выборочных средних (фрагмент рабочего листа MS Excel в режиме отображения данных)

Вычисляем значение $t_{набл}$

$$t_{набл} = \frac{\bar{x} - \bar{y}}{S} = \frac{13.636 - 9.4444}{1.053} = 3.981$$

Вычислим табличное значение $t_{крит}$ с помощью встроенной функции СТЬЮДРАСПОБР(). Для этого определим число степеней свободы по формуле $k = n_1 + n_2 - 2 = 11 + 9 - 2 = 18$, и с учетом уровня значимости α =5% (или α =0,05) получим $t_{крит} = 2{,}100$.

Так как $|t_{набл}| > t_{крит}$, гипотеза H_0 отклоняется, принимается гипотеза H_1. Из этого следует вывод о преимуществе экспериментального обучения.

	B	C
4	x	(x-xср)^2
5	12	=(B5-B18)^2
13	15	=(B13-B18)^2
14	18	=(B14-B18)^2
15	14	=(B15-B18)^2
16	=СУММ(B5:B15)	=СУММ(C5:C15)
17	=СЧЁТ(B5:B15)	
18	=B16/B17	
19	=C16/(B17-1)	
24	=B19/D19	
25		
26		
27		
28	=> можем использовать t-тест с равными диспе	
35	=КОРЕНЬ((C16+E16)/(B17+D17-2)*(1/B17+1/D17))	
36	=(B18-D18)/B35	
37		
38		

Рис. 12.*а.* Проверка гипотезы о совпадении двух выборочных средних (начало) (фрагмент рабочего листа MS Excel в режиме отображения формул)

	D	E
4	y	(y-yср)^2
5	13	=(D5-D18)^2
13	11	=(D13-D18)^2
14		
15		
16	=СУММ(D5:D15)	=СУММ(E5:E15)
17	=СЧЁТ(D5:D15)	
18	=D16/D17	
19	=E16/(D17-1)	
24	alfa=	0.05
25		
26	Fкрит=	=FРАСПОБР(E24;B17-1;D17-1)
27		
28		
35	k=	=(B17+D17-2)
36	alfa=	0.05
38	tкрит=	=СТЬЮДРАСПОБР(E36;E35)

Рис. 12.*б*. Проверка гипотезы о совпадении двух выборочных средних (окончание) (фрагмент рабочего листа Excel в режиме отображения формул)

Средство анализа «Двухвыборочный t-тест с одинаковыми дисперсиями» надстройки «Пакет анализа» MS Excel

Средство анализа «Двухвыборочный *t*-тест с одинаковыми дисперсиями» служит для проверки гипотезы о равенстве средних значений двух независимых нормально распределенных выборок с одинаковыми дисперсиями. Для проверки необходимо заполнить диалоговое окно, приведенное на рис.13. Результат работы представлен на рис.14. Сравните полученные результаты с результатами, полученными вручную.

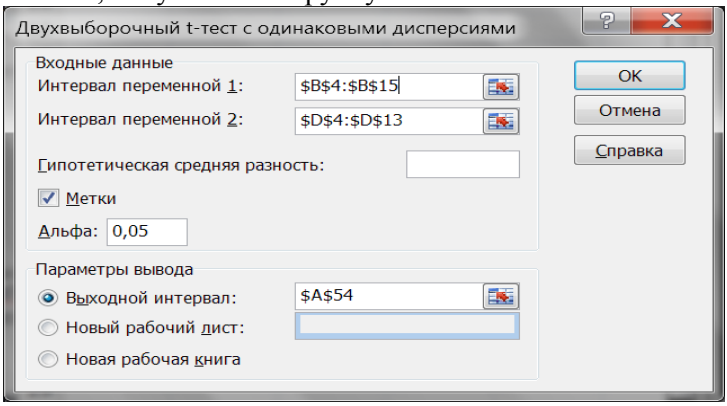


Рис. 13. Диалоговое окно средства анализа «Двухвыборочный *t*-тест с одинаковыми дисперсиями» надстройки «Пакет анализа» MS Excel

	A	B	C
54	**Двухвыборочный t-тест с одинаковыми дисперсиями**		
55			
56		***x***	***y***
57	**Среднее**	**13.6363636**	**9.444444444**
58	**Дисперсия**	**6.05454545**	**4.777777778**
59	**Наблюдения**	**11**	**9**
60	**Объединенная дисперсия**	**5.48709315**	
61	**Гипотетическая разность средни**	**0**	
62	**df**	**18**	
63	**t-статистика**	**3.98147813**	
64	**P(T<=t) одностороннее**	**0.0004376**	
65	**t критическое одностороннее**	**1.73406306**	
66	**P(T<=t) двухстороннее**	**0.0008752**	
67	**t критическое двухстороннее**	**2.10092367**	

Рис. 14. Результат работы средства анализа «Двухвыборочный *t*-тест с одинаковыми дисперсиями» надстройки «Пакет анализа» MS Excel

Рассмотрим второй вариант (дисперсии выборок не равны).

Требуется сравнить средние значения двух независимых выборок, если выборочные дисперсии не равны.

В этом случае значение S вычисляется по формуле:

$$S = \sqrt{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)} \qquad (11)$$

где S_1^2 и S_2^2 - выборочные дисперсии. Число степеней свободы определяется довольно сложным способом. На практике, как правило, оно вычисляется с помощью статистических пакетов в явной или в неявной форме, например, в MS Excel .

Средство анализа «Двухвыборочный t-тест с различными дисперсиями» надстройки «Пакет анализа» MS Excel

Средство анализа «Двухвыборочный *t*-тест с различными дисперсиями» служит для проверки гипотезы о равенстве средних значений двух выборок, взятых из нормально распределенных совокупностей с различными дисперсиями. Для проверки необходимо заполнить диалоговое окно, приведенное на рис.15, назначение всех полей очевидно.

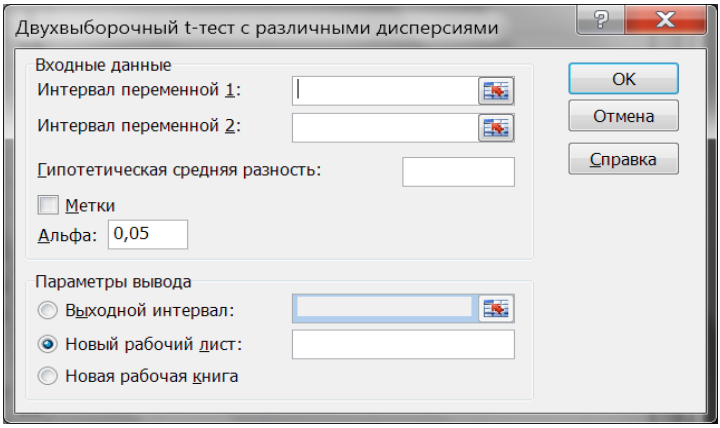


Рис. 15. Диалоговое окно средства анализа «Двухвыборочный t-тест с различными дисперсиями» надстройки «Пакет анализа» MS Excel

3.2. СЛУЧАЙ ДВУХ ЗАВИСИМЫХ ВЫБОРОК

Требуется сравнить средние значения двух зависимых выборок, полученных из нормально распределенной совокупности. Объем выборок одинаков.

В этом случае значения $t_{набл}$ вычисляют по формуле (11), которая в данном случае примет вид:

$$t_{набл} = \frac{\overline{d}}{S_d}, \qquad (12)$$

где $\overline{d} = x_i - y_i$ — разности между соответствующими значениями переменной $\boldsymbol{x}$ и переменной y;

$\overline{d}$ - среднее значение этих разностей;

S_d - стандартная ошибка разности средних значений.

S_d вычисляется по формуле:

$$S_d = \sqrt{\frac{\sum d_i^2 - \frac{\left[\sum d_i\right]^2}{n}}{n \cdot (n-1)}}. \qquad (13)$$

Число степеней свободы k определяется по формуле $k = n - 1$, где n - объем выборки.

Рассмотрим пример использования *t*-критерия Стьюдента для связанных и, очевидно, равных по численности выборок.

Пример 5

Исследовали влияния тренинга на частоту сердечных сокращений (ЧСС) у группы пациентов, страдающих тахикардией. В первом случае пациенты принимали традиционные лекарства, тренинг не проводился, величина ЧСС обозначена через X (рис.16). В другом случае эти же пациенты принимали традиционные лекарства после проведения сеанса тренинга, величина ЧСС обозначена через *Y*. Требуется оценить эффективность проведения сеанса тренинга на ЧСС.

Решение

В столбце **B** (рис.16) содержатся значения $\{x_i\}_{i=1}^{n}$ ЧСС после приема лекарств у пациентов без тренинга, в столбце **C** $\{y_i\}_{i=1}^{n}$ ЧСС при приеме лекарств после сеанса тренинга (n =10). Поскольку группа пациентов одна и та же, в данном примере применима методика для связанных и равных по численности выборок.

Вначале произведем расчет $\overline{d}$ (ячейка D20): $\overline{d} = 2.1$

Затем по формуле (13), получим:

$$S_d = \sqrt{\frac{\sum d_i^{\,2} - \frac{\left[\sum d_i\right]^2}{n}}{n \cdot (n-1)}} = \sqrt{\frac{103 - \frac{[21]^2}{10}}{10 \cdot (10-1)}} = 0.809$$

Далее следует применить формулу (12). Получим:

$$t_{набл} = \frac{\overline{d}}{S_d} = \frac{2.1}{0.809} = 2.596$$

Число степеней свободы: $k = n - 1 = 10 - 1 = 9$. С помощью встроенной функции находим $t_{\text{крит}}$ =СТЬЮДРАСПОБР(2*D23;D22). При вычислении $t_{\text{крит}}$ следует учесть, что в данной задаче следует рассматривать одностороннюю критическую область.

Множитель, равный 2, перед значением уровня значимости добавлен в силу конструктивной особенности этой функции $t_{крит}$ =1,83.

	A	B	C	D	E	F
1	**Критерий Стьюдента (t-критерий)**					
3	**Сравнение двух зависимых выборок**					
4	**Гипотеза H_0 :** средние значения выборок равны					
5	**альтернативная гипотеза H1: среднее Y меньше среднег**					
6						
7	**Пациенты**	**x**	**y**	**d**	**(d)^2**	
8	Иванов	**80**	**76**	**4**	**16**	
9	Новиков	**92**	**92**	**0**	**0**	
10	Сидоров	**80**	**79**	**1**	**1**	
11	Пирогов	**89**	**85**	**4**	**16**	
12	Агапов	**87**	**80**	**7**	**49**	
13	Суворов	**86**	**86**	**0**	**0**	
14	Рыжиков	**85**	**85**	**0**	**0**	
15	Серов	**89**	**90**	**-1**	**1**	
16	Топоров	**90**	**86**	**4**	**16**	
17	Быстров	**92**	**90**	**2**	**4**	
18	**Сумма**	**870**	**849**	**21**	**103**	
19	**Объем выборк**	**10**	**10**	**10**		
20	**Ср.значение**	**87**	**84.9**	**2.1**		
21						
22	**S=**	**0.808977**	**k=**	**9**		
23	**tнабл=**	**2.59587**	**alfa=**	**0.05**		
24						
25	$\lvert t_{набл} \rvert > t_{крит}$		**tкрит=**	**1.83**		
27	**Гипотеза H0 отклоняется, принимаем гипотезу H1.**					

Рис. 16. Проверка гипотезы о совпадении двух выборочных средних в случае двух зависимых выборок

Так как $t_{набл} = 2.596$, то возможно принять альтернативную гипотезу (H_1) о достоверном уменьшении ЧСС у пациентов группы *Y*. Отсюда можно сделать вывод об эффективности тренинга перед приемом лекарств.

В терминах проверки статистических гипотез полученный результат будет звучать так: на 5% уровне гипотеза H_0 отклоняется и принимается гипотеза H_1 .

Средство анализа «Парный двухвыборочный t-тест для средних» надстройки «Пакет анализа» MS Excel

Это средство анализа служит для проверки гипотезы о равенстве средних парных наблюдений, когда наблюдения собраны в пары, и нужно исследовать разницу между ними.

Для проверки необходимо заполнить диалоговое окно, приведенное на рис.17., назначение всех полей очевидно. Результат работы представлен на рис.18. Сравните полученные результаты с результатами, полученными вручную.

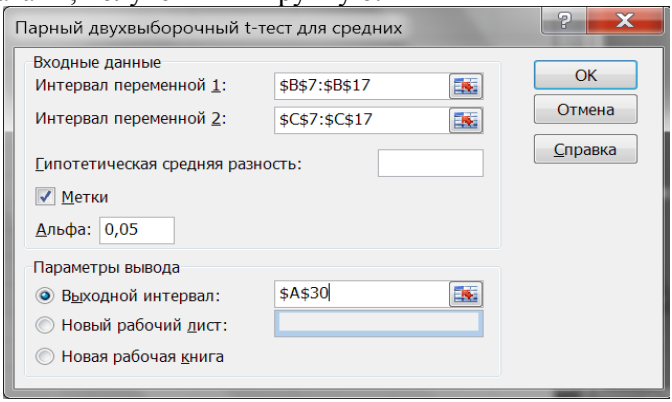


Рис. 17. Диалоговое окно средства анализа «Парный двухвыборочный *t*-тест для средних» надстройки «Пакет анализа» MS Excel

	A	B	C
30	**Парный двухвыборочный t-тест для средних**		
31			
32		***x***	***y***
33	**Среднее**	**87**	**84.9**
34	**Дисперсия**	**18.8888889**	**26.98889**
35	**Наблюдения**	**10**	**10**
36	**Корреляция Пирсона**	**0.87103388**	
37	**Гипотетическая разность средних**	**0**	
38	**df**	**9**	
39	**t-статистика**	**2.59586978**	
40	**P(T<=t) одностороннее**	**0.01446678**	
41	**t критическое одностороннее**	**1.83311386**	
42	**P(T<=t) двухстороннее**	**0.02893356**	
43	**t критическое двухстороннее**	**2.26215889**	

Рис. 18. Результат работы средства анализа «Парный двухвыборочный *t*-тест для средних» надстройки «Пакет анализа» MS Excel

ЗАДАНИЕ

ЗАДАЧА 1

Проверка гипотезы о согласии выборочных данных с нормальным законом распределения. Имеется выборка объема *n* из непрерывно распределенной генеральной совокупности.

Требуется проверить гипотезу, состоящую в том, что выборочные данные получены из нормально распределенной генеральной совокупности.

Варианты заданий взять из лабораторной работы [4].

Решение задачи 1 приведено выше (см. пример 2) .

ЗАДАЧА 2

Текст задачи в зависимости от варианта приведен ниже.

ПРИМЕР ЗАДАЧИ 2

В рабочей книге MS Excel *Лечебницы.xls* содержатся статистические данные, связанные с работой городских и загородных лечебниц, собранные Отделом здравоохранения штата Нью-Мексико.

Фрагмент этой книги приведен на рис.19.

	A	B	C	D	E	F	G
1	**Койки**	**Лечебные дни**	**Всего дней**	**Доход**	**Зарплаты**	**Расходы**	**Расположение**
2	244	128	385	23521	5230	5334	городская
3	59	155	203	9160	2459	493	загородная
51	60	48	213	10644	2820	1154	городская
52	54	119	144	7556	2088	245	загородная
53	120	217	327	20182	4432	6274	городская

Рис. 19. Фрагмент рабочего листа Excel с исходными данными для задачи 2

Выяснить, есть ли разница между доходами городских и загородных лечебниц.

Решение

Выдвигаем гипотезу H_0*:* средние значения доходов городских и загородных лечебниц равны, альтернативная гипотеза H_1 : доходы не равны. Чтобы проверить эту гипотезу с помощью *t*-критерия, необходимо выполнить ряд операций:

1. Разделить всю выборку на две части: для городских и для загородных лечебниц. Считать эти выборки самостоятельными;
2. Выяснить, имеют ли эти выборки одинаковую дисперсию, если «да», то перейти к пункту 3, в противном случае перейти к пункту 4;
3. Применить двухвыборочный *t*-тест с одинаковыми дисперсиями;
4. Применить двухвыборочный *t*-тест с различными дисперсиями.

Пункт 1. Для разделения выборки воспользоваться командой *Данные* → *Сортировка*. Результат приведен на рис.20. В интервале строк 59:76 содержатся данные, относящиеся к городским лечебницам. В интервале строк 77:110 - данные, относящиеся к загородным лечебницам.

	A	B	C	D	E	F	G
58	**Койки**	**Лечебные дни**	**Всего дней**	**Доход**	**Зарплаты**	**Расходы**	**Расположение**
59	244	128	385	23521	5230	5334	городская
74	135	157	471	24274	7485	1344	городская
75	60	48	213	10644	2820	1154	городская
76	120	217	327	20182	4432	6274	городская
77	59	155	203	9160	2459	493	загородная
108	78	154	203	9327	3672	1242	загородная
109	83	224	390	12362	3995	1484	загородная
110	54	119	144	7556	2088	245	загородная

Рис. 20. Фрагмент рабочего листа Excel с данными для задачи 2 после сортировки

Пункт 2. Для проверки предположения, что эти выборки имеют одинаковую дисперсию, воспользуемся критерием Фишера.

Выдвигаем гипотезу H_0: дисперсии выборок равны, альтернативная гипотеза H_1: дисперсии не равны.

Воспользуемся надстройкой MS Excel «Пакет анализа» «Двухвыборочный *F*-тест для дисперсии». Результат расчета приведен на рис.21.

	A	B	C
113	Двухвыборочный F-тест для дисперсии		
114		*Переменная 1*	*Переменная 2*
115	Среднее	16821.55556	12827.61765
116	Дисперсия	50334522.38	43541606.43
117	Наблюдения	18	34
118	df	17	33
119	F	1.156009769	
120	P(F<=f) одностороннее	0.34915504	
121	F критическое одностороннее	1.943000427	

Рис. 21. Фрагмент рабочего листа MS Excel с данными для проверки равенства дисперсий

Поскольку $F_{\text{набл.}} \le F_{\text{крит.}}$ (вычисленное значение критерия $F_{\text{набл.}}$ ***не больше*** критического), то принимается гипотеза H_0 (дисперсии выборок равны). Отсюда следует, что можно применить двухвыборочный *t*-тест с одинаковыми дисперсиями.

Выдвигаем гипотезу H_0: средние арифметические значения выборок равны, альтернативная гипотеза H_1: эти значения не равны.

Воспользуемся надстройкой MS Excel «Пакет анализа» «Двухвыборочный *t*-тест с одинаковыми дисперсиями»; результат работы приведен на рис. 22.

	A	B	C
124	Двухвыборочный t-тест с одинаковыми дисперсиями		
125		*Переменная 1*	*Переменная 2*
126	Среднее	16821.55556	12827.61765
127	Дисперсия	50334522.38	43541606.43
128	Наблюдения	18	34
129	Объединенная дисперсия	45851197.85	
130	Гипотетическая разность средни	0	
131	df	50	
132	t-статистика	2.023485074	
133	P(T<=t) одностороннее	0.024192394	
134	t критическое одностороннее	1.675905423	
135	P(T<=t) двухстороннее	0.048384789	
136	t критическое двухстороннее	2.008559932	

Рис. 22. Фрагмент рабочего листа MS Excel с данными для проверки равенства средних

В качестве $t_{\text{крит}}$ следует рассматривать двустороннее значение. Так как $t_{набл} = 2.02$, $t_{\text{крит}} = 2.008$, то $\left|t_{\text{набл}}\right| > t_{\text{крит}}$,

следовательно, гипотеза H_0 отклоняется, гипотеза H_1 - принимается. Из этого делаем вывод о том, что средние значения доходов городских и загородных лечебниц различны.

ВАРИАНТЫ ЗАДАНИЙ

Вариант 1

В рабочей книге MS Excel *Лечебницы.xls* содержатся статистические данные о работе городских и загородных лечебниц, собранные Отделом здравоохранения штата Нью-Мексико. Определите, есть ли статистически значимая разница между количеством коек в загородных и городских лечебницах.

Вариант 2

В рабочей книге MS Excel *Лечебницы.xls* содержатся статистические данные о работе городских и загородных лечебниц, собранные Отделом здравоохранения штата Нью-Мексико. Верно ли утверждение, что загородные лечебницы используются реже, чем городские?

Указание. В качестве характеристики использования лечебницы ввести переменную «Дней_на_койку», равную отношению количества «Лечебные дни» к значению «Койки» для загородных и городских лечебниц.

Вариант 3

В рабочей книге MS Excel *Лечебницы.xls* содержатся статистические данные о работе городских и загородных лечебниц, собранные Отделом здравоохранения штата Нью-Мексико.

Верно ли утверждение, что загородные лечебницы имеют более низкий объем заработной платы, чем городские?

Вариант 4

В рабочей книге MS Excel *Лечебницы.xls* содержатся статистические данные о работе городских и загородных лечебниц, собранные Отделом здравоохранения штата Нью-Мексико.

Верно ли утверждение, что объем расходов в загородных лечебницах ниже, чем в городских?

Указание. В качестве характеристики объем расходов с учетом разницы в размерах лечебницы ввести переменную «Расход_на_койку», равную отношению количеству «Расходы» к значению «Койки» для загородных и городских лечебниц.

Вариант 5

В рабочей книге MS Excel *ПреподКолледж.xls* содержатся данные о заработной плате преподавателей колледжа. Верно ли утверждение, что преподаватели–женщины получают в среднем меньшую зарплату по сравнению с преподавателями-мужчинами?

Вариант 6

В рабочей книге MS Excel *ПреподКолледж.xls* содержатся данные о заработной плате преподавателей колледжа. Верно ли утверждение, что при поступлении на работу преподаватели со степенью получают в среднем большую зарплату, чем преподаватели без степени?

Вариант 7

В рабочей книге MS Excel *ПреподКолледж.xls* содержатся данные о заработной плате преподавателей колледжа. Верно ли утверждение, что поступающие на работу преподаватели со степенью имеют в среднем больший возраст по сравнению с поступающими на работу преподавателями без степени?

Вариант 8

Рабочая книга MS Excel *ПреподЗатраты.xls* содержит сведения о заработной плате учителей и затратах в общественных школах в пересчете на одного ученика. Верно ли утверждение, что средняя зарплата учителя в северных районах отличается от средней зарплаты учителя в остальных районах ?

Вариант 9

Рабочая книга MS Excel *ПреподЗатраты.xls* содержит сведения о заработной плате учителей и затратах в общественных школах в пересчете на одного ученика. Верно ли утверждение, что средние затраты в общественных школах в пересчете на одного ученика в южных районах отличаются от средних затрат в остальных районах?

Вариант 10

Рабочая книга MS Excel *ПреподЗатраты.xls* содержит данные о заработной плате учителей и затратах в общественных школах в пересчете на одного ученика. Верно ли утверждение, что средняя зарплата учителя в западных районах отличается от средней зарплаты учителя в остальных районах?

Вариант 11

Рабочая книга MS Excel *ПреподЗатраты.xls* содержит данные о заработной плате учителей и затратах в общественных школах в пересчете на одного ученика. Верно ли утверждение, что средние затраты в общественных школах в пересчете на одного ученика в западных районах отличаются от средних затрат в остальных районах?

Вариант 12

В 2007 году распределение спортсменов среди спортивных команд (драфт) было организовано с помощью лотереи: 366 возможных дат рождения спортсменов были помещены во вращающийся барабан и даты были выбраны случайным образом последовательно одна за другой. Первая выбранная дата получила номер 1, вторая – 2 и т.д. В рабочей книге MS Excel *Драфт.xls* содержатся данные о полученных таким образом номерах драфта. Верно ли что утверждение, спортсмены, родившиеся во второй половине года, в среднем имеют более низкие значения драфта, чем спортсмены, родившиеся в первой половине года?

Вариант 13

В 2007 году распределение спортсменов среди спортивных команд (драфт) было организовано с помощью лотереи: 366 возможных дат рождения спортсменов были помещены во вращающийся барабан, и даты были выбраны случайным образом последовательно одна за другой. Первая выбранная дата получила номер 1, вторая – 2 и т.д. В рабочей книге MS Excel *Драфт.xls* содержатся данные о полученных таким образом номерах драфта. Верно ли что утверждение, спортсмены, родившиеся во второй половине месяца в среднем имеют более низкие значения драфта, чем спортсмены, родившиеся в первой половине месяца?

Вариант 14

Рабочая книга MS Excel *Кредиты.xls* содержит данные об отказах в выдаче кредита для 20 кредитных учреждений в зависимости от расы получателя кредита и уровня его дохода. Предполагается, что кредитные учреждения гораздо чаще отказывают представителям национальных меньшинств. Проверьте обоснованность этого утверждения.

Вариант 15

Рабочая книга MS Excel *Кредиты.xls* содержит сведения об отказах в выдаче кредита для 20 кредитных учреждений в зависимости от расы получателя кредита и уровня его дохода. Предполагается, что кредитные учреждения гораздо реже отказывают клиентам с высоким уровнем доходов. Проверьте обоснованность этого утверждения.

Вариант 16

Рабочая книга MS Excel *Кредиты.xls* содержит данные об отказах в выдаче кредита для 20 кредитных учреждений в зависимости от расы получателя кредита и уровня его дохода. Есть ли основание предполагать, что для представителей национальных меньшинств не существует дискриминации?

СПИСОК ЛИТЕРАТУРЫ

1. *Гмурман В.Е.* Теория вероятностей и математическая статистика, изд.9. М.: Высшая школа, 2003, с.480.
2. *Господариков А.П.* Математический практикум. Ч.5. вероятностей и математическая статистика. Теория функций комплексного переменного. Операционное исчисление. Теория поля. СПб: СПГГИ(ТУ), 2003, с.187
3. *Бер К., Кэйри П.* Анализ данных с помощью Microsoft Excel. М.: Вильямс, 2004, с. 560.
4. *Беляев В.В*. Экономико-математические методы и моделирование. Описательная статистика / В.В. Беляев, Т.Р. Косовцева. Методические указания для выполнения лабораторных работ. СПб: РИЦ Горного университета, 2020, с. 35

СОДЕРЖАНИЕ

ЭКОНОМИКО-МАТЕМАТИЧЕСКИЕ МЕТОДЫ И МОДЕЛИРОВАНИЕ

ПРОВЕРКА СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Методические указания к лабораторным работам для студентов бакалавриата направления 21.03.02

Сост. *В.В. Беляев, Т.Р. Косовцева*

Печатается с оригинал-макета, подготовленного кафедрой информатики и компьютерных технологий

Ответственный за выпуск *В.В. Беляев*

Лицензия ИД № 06517 от 09.01.2002

Подписано к печати 03.11.2020. Формат 60×84/16.
Усл. печ. л. 2,3. Усл.кр.-отт. 2,3. Уч.-изд.л. 2,0. Тираж 75 экз. Заказ 820.

Санкт-Петербургский горный университет
РИЦ Санкт-Петербургского горного университета
Адрес университета и РИЦ: 199106 Санкт-Петербург, 21-я линия, 2

ЭКОНОМИКО-МАТЕМАТИЧЕСКИЕ МЕТОДЫ И МОДЕЛИРОВАНИЕ

МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

Методические указания к лабораторным работам для студентов бакалавриата направления 21.03.02

САНКТ-ПЕТЕРБУРГ
2020

Министерство науки и высшего образования Российской Федерации

Федеральное государственное бюджетное образовательное
учреждение высшего образования
Санкт-Петербургский горный университет

Кафедра информатики и компьютерных технологий

ЭКОНОМИКО-МАТЕМАТИЧЕСКИЕ МЕТОДЫ И МОДЕЛИРОВАНИЕ

МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

Методические указания к лабораторным работам
для студентов бакалавриата направления 21.03.02

САНКТ-ПЕТЕРБУРГ
2020

УДК 519.86:622.3.012 (073)

ЭКОНОМИКО-МАТЕМАТИЧЕСКИЕ МЕТОДЫ И МОДЕЛИРОВАНИЕ. МНОЖЕСТВЕННАЯ РЕГРЕССИЯ. Методические указания к лабораторным работам / Санкт-Петербургский горный университет. Сост.: *В.В. Беляев, Т.Р. Косовцева*. СПб, 2020. 42 с.

Методические указания содержат сведения, необходимые для лабораторных работ по проведению корреляционно-регрессионного анализа. Приведены необходимые теоретические сведения и примеры выполнения заданий по исследованию корреляционных и регрессионных связей между характеристиками экономических процессов, которые являются теоретической основой построения эконометрических моделей. Все решения выполнены с использованием электронных таблиц MS Excel, в том числе с применением надстройки «Пакет анализа».

Предназначены для студентов бакалавриата направления 21.03.02 «Землеустройство и кадастры» дневной формы обучения.

Научный редактор доц. *А.Б. Маховиков*

Рецензент канд. техн. наук *К.В. Столяров* (Корпорация «Телум Инж»)

ВВЕДЕНИЕ

Как правило, реальные экономические явления достаточно сложны и выявление характера связи между различными свойствами (параметрами) таких явлений является сложной задачей. Парная регрессия, рассмотренная в предыдущих лабораторных работах, описывает исследуемую характеристику экономического явления (отклик) в зависимости от одной объясняющей характеристики (фактора) в предположении, что влиянием других факторов можно пренебречь. Адекватное уравнение в этом случае удается построить далеко не всегда, поскольку причиной изменения отклика является одновременное воздействие множества факторов. Для того, чтобы учесть это воздействие необходимо использовать *модель множественной регрессии.*

Построение модели множественной регрессии включает несколько этапов:

- выбор формы связи (уравнения регрессии);
- отбор факторных признаков.

Выбор формы связи затрудняется тем, что, теоретическая зависимость между признаками может быть выражена большим числом различных функций. Поскольку уравнение регрессии строится главным образом для объяснения и количественного отображения взаимосвязей, оно должно хорошо отражать сложившиеся между откликом и исследуемыми факторами фактические связи.

В данной работе описан математический аппарат для построения линейного уравнения множественной регрессии.

ЛАБОРАТОРНАЯ РАБОТА. МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

Цель: освоить на практике нахождение с помощью табличного процессора MS Excel числовых характеристик множественной регрессии, а также изучить основные свойства теории корреляции.

1. ТЕОРЕТИЧЕСКИЕ СВЕДЕНИЯ. БАЗОВЫЕ ПОНЯТИЯ

Будем предполагать, что несколько переменных $\boldsymbol{X_1, X_2, ..., X_p}$ (объясняющих переменных, предикторов, факторных признаков, регрессоров) оказывают воздействие на значения зависимой переменной $\boldsymbol{Y}$ (отклик, результативный признак.

В этом случае целесообразно строить уравнение множественной регрессии.

Множественная регрессия – уравнение связи зависимой переменной Y с независимыми переменными $X_1, X_2, ..., X_p$:

$$Y = f\left(X_1, X_2, ..., X_p\right)$$

Наиболее простой и самой употребляемой является *модель множественной линейной регрессии,* которая имеет вид

$$Y = b_0 + b_1 X_1 + b_2 X_2 + ... + b_p X_p \qquad (1)$$

где $b_0, b_1, ..., b_p$ - параметры уравнения.

Пусть имеется *n*-наблюдений, тогда исходные данные представимы в виде матрицы размерности *n* на *p* и вектора размерности *n* :

$$\begin{bmatrix} X_1^1 & X_2^1 & ... & X_p^1 \\ X_1^2 & X_2^2 & ... & X_p^2 \\ ... & ... & ... & ... \\ X_1^n & X_2^n & ... & X_p^n \end{bmatrix}, \begin{bmatrix} Y_1 \\ . \\ . \\ . \\ Y_n \end{bmatrix}. \qquad (2)$$

Все элементы *i*-ой строки $\left(X_1^i \quad X_2^i \quad ... \quad X_p^i\right)$ и *i*-ого элемента вектора Y_i - результаты *i*-ого наблюдения. Будем предполагать, что все наблюдения независимы и получены примерно в одинаковых условиях. В этом случае набор данных, определяемый соотношениями (2) называют *пространственной выборкой или пространственными данными (cross section data)*.На практике эти значения часто получаются как результаты некоторого эксперимента, поэтому их часто называют наблюдаемыми или экспериментальными или эмпирическими значениями.

Для оценки параметров уравнения множественной регрессии применяют метод наименьших квадратов (МНК). Идея этого метода была подробно рассмотрена в лабораторной работе «Линейная парная регрессия»[1]. Все соображения и выводы применимы и в случае множественной линейной регрессии с поправкой на количество факторов.

Рассчитаем $\hat{Y}_i$ - теоретические значения отклика, подставив в уравнение (1) значений факторных переменных *i*-го наблюдения. В результате получим величину

$$\hat{Y}_i = b_0 + b_1 X_1^i + b_2 X_2^i + ... + b_p X_p^i , \qquad (3)$$

Значения $\hat{Y}_i$ практически никогда не будут совпадать с наблюдаемыми значениями Y_i .

Разность между наблюдаемыми значениями Y_i и значениями $\hat{Y}_i$, рассчитанным по уравнению регрессии, называется *регрессионным остатком* в наблюдении *i* и обозначается ε_i :

$$\varepsilon_i = Y_i - \hat{Y}_i . \quad i = \overline{1,n} , \qquad (4)$$

Отметим, что ε_i , $i = \overline{1,n}$ являются случайными величинами, которые также называют *случайными компонентами, случайными членами, возмущениями или остатками*.

С учетом соотношения (4), справедливо соотношение

$$Y_i = \hat{Y}_i + \varepsilon_i = b_0 + b_1 X_1^i + b_2 X_2^i + ... + b_p X_p^i + \varepsilon_i . \qquad (5)$$

Присутствие в этом соотношении случайной компоненты ε_i, обусловлено следующими причинами:

- ошибками спецификации, то есть отбора факторов, и выбора связи между явлениями;
- ошибками измерения.

Будем полагать, что относительно ε выполняется ряд утверждений, известных как *условия Гаусса-Маркова:*

1. Равенство нулю математического ожидания регрессионных остатков:
   $M(\varepsilon_i) = 0,\ i = 1,...,n$; (6)
2. Постоянство дисперсии регрессионных остатков (гомоскедастичность остатков):
   $M(\varepsilon_i^{\,2}) = D(\varepsilon_i) = \sigma^2$; (7)
3. Отсутствие систематической связи (корреляции) между значениями регрессионных остатков в любых двух наблюдениях: $M(\varepsilon_i \cdot \varepsilon_j) = 0 \quad (i \neq j)$; (8)
4. $X_1, X_2,..., X_p$ - неслучайные величины.

Для определения параметров $b_0, b_1,..., b_p$ уравнения множественной линейной регрессии по МНК составляется сумма $S_{\text{ост}}$ - *остаточная сумма квадратов*

$$S_{\text{ост}} = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i\,x_1 x_2 \ldots x_p})^2 . \quad (9)$$

Она равна сумме квадратов отклонений (остатков) наблюдаемых (эмпирических) значений отклика Y_i от теоретических значений $\hat{Y}_i$ в точке X_i . Чтобы подчеркнуть её зависимость от параметров уравнения регрессии $b_0, b_1,..., b_p$, обозначим её как функцию от этих параметров через $S(b_0, b_1,..., b_p)$.

$$S(b_0, b_1,..., b_p) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (Y_i - \hat{Y}_i)^2 . \quad (10)$$

Цель метода наименьших квадратов (МНК) заключается в выборе таких оценок $b_0, b_1, ..., b_p$ *, для которых сумма квадратов отклонений (остатков) будет минимальной.*

Для того чтобы найти набор коэффициентов $b_0, b_1, ..., b_p$, которые доставляют минимум функции $S(b_0, b_1, ..., b_p)$, используем необходимое условие экстремума функции нескольких переменных - равенство нулю частных производных

$$\frac{\partial S(b_0, b_1, ..., b_p)}{\partial b_0} = 0; \quad \frac{\partial S(b_0, b_1, ..., b_p)}{\partial b_1} = 0; \quad ... \frac{\partial S(b_0, b_1, ..., b_p)}{\partial b_p} = 0$$

В результате преобразований получаем следующую систему нормальных уравнений:

$$\begin{cases} \sum Y = n \cdot b_0 + b_1 \cdot \sum X_1 + b_2 \cdot \sum X_2 + ... + b_p \cdot \sum X_p \\ \sum YX_1 = b_0 \sum X_1 + b_1 \cdot \sum X_1^2 + b_2 \cdot \sum X_1 X_2 + ... + b_p \cdot \sum X_p X_1 \\ ..... \\ \sum YX_p = b_0 \sum X_p + b_1 \cdot \sum X_1 X_p + b_2 \cdot \sum X_p X_2 + ... + b_p \cdot \sum X_p^2 \end{cases} \qquad (11)$$

Для ее решения может быть применен любой известный метод решения системы линейных уравнений.

Коэффициенты $\{b_j\}_{j=1}^{p}$ в уравнении (3) называются *коэффициентами множественной регрессии*. Величина коэффициента b_j показывает среднее изменение отклика *Y* при изменении фактора $\boldsymbol{X_j}$ на единицу.

Другой вид уравнения множественной регрессии - уравнение регрессии в стандартизованном масштабе:

$$t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2} + ... + \beta_p t_{x_p}, \qquad (12)$$

где: $t_y = \frac{Y - \overline{Y}}{\sigma_y}, t_{x_i} = \frac{X_i - \overline{X}_i}{\sigma_{x_i}}$ - стандартизованные переменные;

$i = 1,.., p$, $\boldsymbol{p}$ - число неизвестных;

$\overline{Y}$, $\overline{X}_i$ - средние значения;

σ_y, σ_{x_i} - средние квадратические отклонения;

β_i -стандартизованные коэффициенты регрессии.

В силу того, что стандартизованные переменные заданы как центрированные (средние значения $\bar{t}_y = \bar{t}_x = 0$) и нормированные (средние квадратические отклонения $\sigma_{t_y} = \sigma_{t_x} = 1$), стандартизованные коэффициенты регрессии сравнимы между собой, и с их помощью можно ранжировать факторы по силе их воздействия на результат.

Для определения коэффициентов уравнения множественной регрессии в стандартизованном масштабе так же применим МНК. Коэффициенты $\{\beta_j\}_{j=1}^{p}$ можно получить, решая систему, аналогичную системе (7). Эту систему можно преобразовать, и тогда, стандартизованные коэффициенты регрессии определяются из следующей системы уравнений:

$$\begin{cases} r_{yx_1} = \beta_1 + \beta_2 r_{x_2x_1} + \beta_3 r_{x_3x_1} + ... + \beta_p r_{x_Px_1} \\ r_{yx_2} = \beta_1 r_{x_2x_1} + \beta_2 + \beta_3 r_{x_3x_2} + ... + \beta_p r_{x_px_2} \\ ..... \\ r_{yx_p} = \beta_1 r_{x_px_1} + \beta_2 r_{x_px_2} + \beta_3 r_{x_3x_p} + ... + \beta_p \end{cases}, \qquad (13)$$

где

- $r_{x_ix_j}$ - коэффициент парной корреляции между факторами X_i и X_j,
- r_{yx_j} - коэффициент парной корреляции между откликом Y и фактором X_j .

Отметим, что связь коэффициентов множественной регрессии b_j со стандартизованными коэффициентами β_j описывается соотношением

$$b_j = \beta_j \frac{\sigma_y}{\sigma_{x_j}}. \qquad (14)$$

Стандартизованный коэффициент регрессии β_j показывает, на сколько величин σ_y в среднем изменится отклик при увеличении *j-го* фактора на одну величину σ_{x_j} .

Средние коэффициенты эластичности для линейной регрессии рассчитываются по формуле :

$$\overline{Э}_{yx_j} = b_j \frac{\overline{X}_j}{\overline{Y}}. \qquad (15)$$

Средний коэффициент эластичности $\overline{Э}_{yx_j}$ показывает на сколько процентов в среднем изменится отклик при изменении его среднего значения фактора X_j на один процент, при неизменном значении остальных факторов.

2. КОЭФФИЦИЕНТ ДЕТЕРМИНАЦИИ

Тесноту совместного влияния факторов на результат показывает *множественной детерминации.*

Качество построенной модели в целом оценивается коэффициентом множественной детерминации, который определяется формулой:

$$R^2_{yx_1x_2\ldots x_p} = 1 - \frac{S_{\text{ост}}}{S_{\text{полн}}}, \qquad (16)$$

где $S_{\text{ост}} = \sum_{i=1}^{n} (Y_i - \hat{Y}_{i\,x_1x_2\ldots x_p})^2$ - *остаточная сумма квадратов отклонений*,

$S_{\text{полн}} = \sum_{i=1}^{n} (Y_i - \overline{Y})^2$ - общая сумма квадратов отклонений значений Y_i от среднего арифметического значения отклика Y .

Для линейной регрессии справедливо следующее равенство:

$S_{\text{полн}} = S_{\text{ост}} + S_{\text{регр}}$,

где $S_{\text{регр}} = \sum_{i=1}^{n} (\hat{Y}_{i x_1 x_2 \ldots x_p} - \overline{Y})^2$ *регрессионная сумма квадратов отклонений*.

Остаточная сумма квадратов отклонений $S_{\text{ост}}$ характеризует суммарное отклонение наблюдаемых (эмпирических) данных от теоретических значений, найденных по уравнению регрессии. Регрессионная или факторная сумма квадратов отклонений $S_{\text{регр}}$ характеризует разброс теоретических значений относительно среднего арифметического значения наблюдаемого значения (отклика).

Все свойства коэффициента детерминации $R^2_{yx_1x_2\ldots x_p}$ указаны в лабораторной работе [1]. Так, значение этого коэффициента лежит в пределах от нуля до единицы. Это значение показывает долю объясненной вариации результативного признака (отклика) за счет включенных в уравнение p факторов, т.е. насколько хорошо уравнение, полученное с помощью регрессионного анализа, объясняет взаимосвязь между откликом и факторами. Доля необъясненной вариации отклика других, не учтенных в модели факторов, равна $1 - R^2$. Коэффициент детерминированности служит показателем тесноты связи между независимой переменной и факторами. Показателю тесноты связи можно дать качественную оценку (шкала Чеддока)

Таблица 1

Шкала Чеддока

Количественная мера тесноты связи	***Качественная характеристика силы связи***
0,1-0,3	Слабая
0,3-0,5	Умеренная
0,5-0,7	Заметная
0,7-0,9	Высокая
0,9-0,99	Весьма высокая

Величину R^2 для уравнения множественной регрессии в стандартизованном масштабе можно определить по формуле

$$R^2 = \sum \beta_i \cdot r_{yx_i} \qquad (17)$$

3. ОЦЕНКА НАДЕЖНОСТИ УРАВНЕНИЯ МНОЖЕСТВЕННОЙ РЕГРЕССИИ

Проверить значимость уравнения регрессии - значит установить насколько хорошо математическая модель, выражающая зависимость отклика от факторов, согласуется с экспериментальными данными, с учетом количества наблюдений и количества факторов в уравнении. Оценка значимости уравнения регрессии в целом сводится к проверке того, что величина R^2 не случайно отлична от нуля. Для оценки значимости уравнения множественной регрессии в целом используется ***F***-критерий Фишера.

Выдвигаем нулевую гипотезу $H_0 : R^2 = 0$. Это возможно, когда уравнение регрессии незначимо, т.е. связь между откликом и факторами отсутствует. Альтернативная гипотеза $H_1 : R^2 > 0$, в этом случае уравнение регрессии адекватно описывает связь между откликом и факторами.

Схема проведения дисперсионного анализа приведена в табл.2. Схемы применения ***F***-критерия Фишера для оценки значимости уравнения множественной регрессии и уравнения парной регрессии одинаковы. Различие состоит только в одном – в определении числа степеней свободы $df_{\text{регр}}$ и $df_{\text{ост}}$.

Существует соотношение между числом *степеней свободы* df (числом свободы независимого варьирования признака) для общей, факторной и остаточной сумм квадратов:

$$df_{\text{полн}} = df_{\text{ост}} + df_{\text{регр}} .$$

Для множественной линейной регрессии:

$$df_{\text{полн}} = n - 1;\ df_{\text{регр}} = p;\ df_{\text{ост}} = n - p - 1 ,$$

где n - число единиц совокупности, p - число факторов, включенных в уравнение регрессии.

$$D_{\text{полн}} = \frac{S_{\text{полн}}}{df_{полн}} = \frac{\sum_{i=1}^{n}\left(Y_i - \overline{Y}\right)^2}{n-1}; \qquad (14)$$

$$D_{\text{регр}} = \frac{S_{\text{регр}}}{df_{регр}} = \frac{\sum_{i=1}^{n}\left(\hat{Y}_i - \overline{Y}\right)^2}{p}; \qquad (15)$$

$$D_{\text{ост}} = \frac{S_{\text{ост}}}{df_{\text{ост}}} = \frac{\sum_{i=1}^{n}\left(Y_i - \hat{Y}_i\right)^2}{n-p-1}. \qquad (16)$$

Таблица 2

Схема проведения дисперсионного анализа

Источники вариации:	***Вариация, объясненная за счет регрессии***	***Остаточная вариация***	***Общая вариация***
Число степеней свободы	$df_{\text{регр}} = p$ (p - количество факторов)	$df_{\text{ост}} = n - p - 1$	$df_{\text{полн}} = n - 1$
Сумма квадратов отклонений	$S_{\text{регр}} = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2$	$S_{\text{ост}} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2$	$S_{\text{полн}} = \sum_{i=1}^{n}(y_i - \overline{y})^2$
Дисперсия на одну степень свободы	$D_{\text{регр}} = \frac{S_{\text{регр}}}{df_{\text{регр}}}$	$D_{\text{ост}} = \frac{S_{\text{ост}}}{df_{\text{ост}}}$	$D_{\text{полн}} = \frac{S_{\text{ост}}}{df_{\text{полн}}}$
Фактическое значение критерия Фишера	$F_{\text{набл}} = \frac{D_{\text{регр}}}{D_{\text{ост}}}$		
Табличное значение критерия Фишера	$F_{\text{крит}}$ определяется по уровню значимости и числу степеней свободы числителя $df_{регр}$ и знаменателя $df_{ост}$		

Критерий Фишера определяется следующим соотношением:

$$F_{\text{набл}} = \frac{D_{\text{регр}}}{D_{\text{ост}}} \qquad (17)$$

Использование критерия Фишера предполагает вычисление $F_{\text{набл}}$ и его сравнение с табличным значением $F_{\text{крит}}$, которое зависит от уровня значимости $\boldsymbol{\alpha}$ и числа степеней свободы для регрессионной и остаточной сумм. $F_{\text{крит}}$ определяется либо с помощью таблиц, либо с использованием специализированных пакетов программ, например, в MS Excel для этого может быть использована функция **FРАСПРОБР().**

Если $F_{\text{набл}} > F_{\text{крит}}$, нулевая гипотеза H_0 об отсутствии связи признаков отклоняется и делается вывод о справедливости гипотезы H_1 (о существенности этой связи, значимости уравнения регрессии). Если же величина $F_{\text{набл}}$ окажется меньше табличной, то есть $F_{\text{набл}} < F_{\text{крит}}$, то вероятность нулевой гипотезы H_0 выше заданного уровня значимости (например, **0.05**) и гипотеза H_0 не может быть отклонена без серьезного риска сделать неправильный вывод о наличии линейной связи между факторами X_1 ,X_2, ... ,X_p и откликом ***Y*** . В этом случае уравнение регрессии считается статистически незначимым, и это уравнение нельзя использовать для анализа и прогноза.

Значение $F_{\text{набл}}$ может быть вычислено как по формуле (17), так и с помощью коэффициента детерминированности по формуле (18).

$$F_{\text{набл}} = \frac{D_{\text{регр}}}{D_{\text{ост}}} = \frac{R^2}{1-R^2} \cdot \frac{n-p-1}{p}, \qquad (18)$$

Частный ***F***-критерий оценивает статистическую значимость каждого из факторов в уравнении. В общем виде для фактора $\boldsymbol{X_i}$ частный ***F***-критерий определится как

$$F_{x_i} = \frac{R^2_{yx_1x_2\ldots x_i\ldots x_p} - R^2_{yx_1x_2\ldots x_{i-1}x_{i+1}\ldots x_p}}{1 - R^2_{yx_1x_2\ldots x_i\ldots x_p}} \cdot \frac{n-p-1}{1}. \qquad (19)$$

Необходимость вычисления такой оценки обусловлена тем, что не каждый внесенный в модель фактор будет увеличивать долю объясненной вариации результативного признака. Также при наличии в модели нескольких факторов они могут вводиться в модель в разной последовательности. Ввиду корреляции между факторами значимость одного и того же фактора может быть различной в зависимости от последовательности введения его в модель.

Частный *F*-критерий, вычисленный по формуле (19) построен на сравнении прироста факторной дисперсии, обусловленного влиянием дополнительного включенного фактора (числитель) с остаточной дисперсией на одну степень свободы по регрессионной модели в целом (знаменатель).

Вычисленный частный *F*-критерий F_{x_i} сравнивается с табличным значением $F_{\text{крит}}$, который зависит от уровня значимости $\boldsymbol{\alpha}$ и числа степеней свободы: 1 и $(n-p-1)$. $F_{\text{крит}}$ определяется либо с помощью таблиц, либо с использованием функций, например, функции MS Excel **FРАСПРОБР().**

Если $F_{x_i} > F_{\text{крит}}$, то дополнительное включение фактора $\boldsymbol{X_i}$ в модель статистически оправдано и коэффициент регрессии $\boldsymbol{b_i}$ при $\boldsymbol{X_i}$ статически значим. Если же величина $\boldsymbol{F_{x_i}}$ окажется меньше табличной, то есть $F_{x_i} < F_{\text{крит}}$, то дополнительное включение в модель фактора $\boldsymbol{X_i}$ статистически не оправдано, поскольку существенно не увеличивает долю объясненной вариации отклика. При этом коэффициент регрессии $\boldsymbol{b_i}$ при $\boldsymbol{X_i}$ статистически незначим, что еще раз подтверждает нецелесообразность включения этого фактора в модель.

С частным *F*-критерием тесно связан $\boldsymbol{t}$ -критерий Стьюдента для проверки значимости коэффициентов регрессии.

Оценка значимости коэффициентов множественной регрессии с помощью ***t*** -критерия Стьюдента производится с помощью величины t_{b_i} , вычисляемой по формулам

$$t_{b_i} = \frac{b_i}{m_{b_i}}, \qquad (20)$$

или

$$\left|t_{b_i}\right| = \sqrt{F_{x_i}} \qquad (21)$$

где m_{b_i} – средняя квадратическая ошибка коэффициента регрессии $\boldsymbol{b_i}$, которая может быть определена по следующей формуле:

$$m_{b_i} = \frac{\sigma_y \cdot \sqrt{1 - R^2_{yx_{1\ldots}x_p}}}{\sigma_{x_i} \cdot \sqrt{1 - R^2_{x_i x_{1\ldots} x_p}}} \cdot \frac{1}{\sqrt{n - p - 1}} \quad , \qquad (22)$$

где $R^2_{x_i x_{1\ldots} x_p}$ - коэффициент множественной детерминации для зависимости фактора $\boldsymbol{X_i}$ от всех других факторов уравнения множественной детерминации, все остальные обозначения очевидны.

Заметим, что формула (21) полностью аналогична формуле (5.18)[1]. Сама процедура проверки значимости соответствующего коэффициента полностью аналогична процедуре проверки в лабораторной работе [1].

При представлении результатов множественной регрессии наряду с уравнением и скорректированным коэффициентом множественной детерминации принято приводить значения t_{b_i} . На практике если наблюдаемые значения $t_{b_i} > 3$, то это означает, что значение этого коэффициента статистически достоверно, а уравнение может быть использовано для прогнозирования.

При эконометрическом исследовании необходимо стремиться к увеличению числа наблюдений, так как большой объем наблюдений является одной из предпосылок признания значимым как уравнения регрессии, так и его коэффициентов. Значимость этих

величин является необходимым условием построения адекватных статистических моделей.

Как показывает практика, для того, чтобы уравнение было адекватным, необходимо, чтобы количество наблюдений n превышало количество определяемых коэффициентов регрессии p в 6-7 раз.

4. СКОРРЕКТИРОВАННЫЙ ИНДЕКС МНОЖЕСТВЕННОЙ ДЕТЕРМИНАЦИИ

Скорректированный (исправленный, adjustable) коэффициент множественной детерминации R^2_{adj} содержит поправку на число степеней свободы и рассчитывается по формуле :

$$R^2_{adj} = 1 - (1 - R^2) \cdot \frac{(n-1)}{(n-p-1)} \qquad (23)$$

Скорректированный коэффициент множественной детерминации R^2_{adj} используется для сопоставления моделей содержащих различное количество факторов.

Чем больше ***p***, тем больше различие между R^2_{adj} и R^2. Чем больше объем выборки ***n***, тем меньше это различие.

Существенно различным может быть изменение R^2 и R^2_{adj} при включении дополнительного фактора в уравнение регрессии. Если этот фактор существенно влияет на отклик, то увеличатся значения как R^2 так и R^2_{adj}. Если вновь добавленный фактор несущественно влияет на отклик, то значение R^2, как правило, увеличивается (может быть незначительно), а значение R^2_{adj} - уменьшается. Очевидно, что в этом случае такой фактор в уравнение включать не целесообразно.

5. ЧАСТНАЯ КОРРЕЛЯЦИЯ

Частные коэффициенты (или индексы) корреляции, измеряющие влияние на Y фактора X_i при устранении влияния

других факторов, включенных в уравнение регрессии, можно определить по формуле:

$$r_{yx_1x_2\ldots x_{i-1}x_{i+1}\ldots x_p} = \sqrt{1-\frac{1-R^2_{yx_1x_2\ldots x_i\ldots x_p}}{1-R^2_{yx_1x_2\ldots x_{i-1}x_{i+1}\ldots x_p}}} \quad , \qquad (24)$$

где $R^2_{yx_1x_2\ldots x_i\ldots x_p}$ - коэффициент детерминированности для уравнения регрессии, в которое включены факторы $X_1, X_2, \ldots, X_p$; $R^2_{yx_1x_2\ldots x_{i-1}x_{i+1}\ldots x_p}$ - коэффициент детерминированности для уравнения регрессии, в которое включены факторы $X_1, X_2, \ldots, X_{i-1}, X_{i+1}, \ldots X_p$, т.е. фактор X_i исключен из уравнения.

Частные коэффициенты корреляции изменяются в пределах от **0** до **1**.

Нетрудно показать, что величина, стоящая под радикалом в правой части равенства (24) может быть преобразована к следующему виду:

$$1-\frac{1-R^2_{yx_1x_2\ldots x_i\ldots x_p}}{1-R^2_{yx_1x_2\ldots x_{i-1}x_{i+1}\ldots x_p}} = \frac{R^2_{yx_1x_2\ldots x_i\ldots x_p}-R^2_{yx_1x_2\ldots x_{i-1}x_{i+1}\ldots x_p}}{1-R^2_{yx_1x_2\ldots x_{i-1}x_{i+1}\ldots x_p}}$$

Правая часть последнего равенства представляет собой отношение приращения объясненной часть вариации отклика за счет включения фактора X_i в уравнение регрессии к необъясненной доле вариации отклика, имевшей место до введения фактора X_i в уравнение регрессии.

Таким образом, величина $r_{yx_i\cdot x_1x_2\ldots x_p}$ характеризует возрастание коэффициента детерминации за счет введения в уравнение регрессии фактора X_i . Благодаря этому частные коэффициенты корреляции могут быть использованы для ранжирования влияния факторов на результат.

Так, при двух факторах и *i*=1 частный коэффициент корреляции $r_{yx1\cdot x2}$ может быть вычислен по формуле

$$r_{yx_1 \cdot x_2} = \frac{r_{yx_1} - r_{yx_2} \cdot r_{x_1 x_2}}{\sqrt{(1 - r_{yx_2}{}^2) \cdot (1 - r_{x_1 x_2}{}^2)}} \qquad (25)$$

Коэффициент $r_{yx_1 \cdot x_2}$ показывает тесноту связи между Y и X_1 при неизменном уровне фактора X_2, включенного в уравнение регрессии.

Аналогично $r_{yx_2 \cdot x_1}$ можно определить по формуле

$$r_{yx_2 \cdot x_1} = \frac{r_{yx_2} - r_{yx_1} \cdot r_{x_1 x_2}}{\sqrt{(1 - r_{yx_1}{}^2) \cdot (1 - r_{x_1 x_2}{}^2)}} \qquad (26)$$

Коэффициенты частной корреляции используют для оценки целесообразности включения фактора в уравнение регрессии.

6. МАТРИЧНАЯ ФОРМА ЗАПИСИ

Матричная форма записи для определения коэффициентов множественной линейной регрессии полностью аналогична таковой для парной регрессии (5.28) [1] , т.е.

$$B = (X^{\mathrm{T}} \cdot X)^{-1} \cdot X^{\mathrm{T}} \cdot y \qquad (27)$$

где $\boldsymbol{X}$ матрица размерности $n \cdot (p+1)$, В – вектор коэффициентов размерности $(p+1)$

$$X = \begin{bmatrix} 1 & X_{11} & . & X_{1p} \\ . & . & . & . \\ . & . & . & . \\ . & . & . & . \\ 1 & X_{n1} & . & X_{np} \end{bmatrix}, \; y = \begin{bmatrix} Y_1 \\ . \\ . \\ . \\ Y_n \end{bmatrix}, \; B = \begin{bmatrix} b_0 \\ b_1 \\ . \\ b_{\mathrm{p}} \end{bmatrix}$$

7. МУЛЬТИКОЛЛИНЕАРНОСТЬ ФАКТОРОВ

При построении уравнения множественной регрессии может возникнуть проблема *мультиколлинеарности* факторов.

Одним из условий построения корректной регрессионной модели является условие линейной независимости факторов. Если это условие нарушается, т.е. *если один из факторов может быть выражен через несколько других, то говорят что, существует полная коллинеарность*.

Это порождает множество проблем. Например, применение формулы (18) невозможно, поскольку матрица $(X^{\mathrm{T}} \cdot X)^{-1}$ не может быть вычислена (определитель $\det(X^{\mathrm{T}} \cdot X) = 0$).

На практике полная коллинеарность встречается редко, гораздо чаще встречается ситуация, когда между факторами наблюдается высокая степень корреляции, и тогда говорят о наличии *мультиколлинеарности* факторов.

В этом случае применение формулы (18) формально возможно, поскольку матрица $(X^{\mathrm{T}} \cdot X)^{-1}$ может быть вычислена (определитель $\det(X^{\mathrm{T}} \cdot X) \neq 0$, но близок к нулю), поэтому полученные значения найденных коэффициентов будут обладать «плохими свойствами».

Основные отрицательные проявления мультиколлинеарности заключаются в следующем:

- Значения найденных коэффициентов модели имеют неправильные с точки зрения теории знаки или неоправданно большие (маленькие) значения.
- Небольшие изменения исходных данных приводит к существенному изменению найденных коэффициентов модели
- Оценки имеют большие стандартные ошибки, малую значимость (хотя вся модель в целом является значимой).
- Невозможно оценить воздействие на отклик каждого фактора в отдельности.

Когда два фактора сильно коррелированы, говорят о коллинеарности факторов.

Считается, что *два фактора* явно *коллинеарны*, т.е. находятся между собой в линейной зависимости, если

$$\left| r_{x_i x_j} \right| \geq 0{,}7. \qquad (28)$$

Для решения проблемы мультиколлинеарности зависимые факторы исключают из модели.

В случае двух явно коллинеарных факторов уравнения регрессии рекомендуется один исключить. Предпочтение при этом отдается тому фактору, который при достаточно тесной связи с результатом имеет наименьшую тесноту связи с другими факторами.

8. ПОСТРОЕНИЕ ПРОГНОЗА С ПОМОЩЬЮ УРАВНЕНИЯ РЕГРЕССИИ

Рассмотрим применение уравнения регрессии для построения точечного и интервального прогноза. Точечный прогноз $\hat{Y}$ может быть получен путем подстановки в уравнение регрессии (1) значений факторов.

Результат точечного прогноза маловероятен. Поэтому находят интервальную оценку прогноза.

Для получения интервальной оценки необходимо воспользоваться формулой аналогичной формуле парной регрессии, которая для множественной регрессии имеет следующий вид.

$$\hat{Y} - \varepsilon \leq \widetilde{Y} \leq \hat{Y} + \varepsilon \qquad (29)$$

где ε - полуширина доверительного интервала.

Точечное значение $\hat{Y}$ является серединой доверительного интервала, $\left(\hat{Y} - \varepsilon\right)$ - левой границей, $\left(\hat{Y} + \varepsilon\right)$ - правой границей.

Величина равна ε половине ширины доверительного интервала и может быть вычислена по формуле

$$\varepsilon = t_{\text{крит}} \cdot s_{\hat{y}} \qquad (30)$$

где $t_{\text{крит}}$ - критическое значение распределения Стьюдента с числом степеней свободы равным n-p-1;

$s_{\hat{y}}$ - стандартная ошибка групповой средней

$$s_{\hat{y}} = s \cdot \sqrt{(X0)^T \left(X^T X\right) X0} \qquad (31)$$

$$X0 = \begin{pmatrix} 1 \\ x_{1,0} \\ x_{2,0} \\ \dots \\ x_{p,0} \end{pmatrix}, \; X = \begin{bmatrix} 1 & X_{11} . & X_{1\,p} \\ . & . \; . & . \\ . & . \; . & . \\ . & . \; . & . \\ 1 & X_{n1} . & X_{n\,p} \end{bmatrix}, \qquad (32)$$

$X0$ – вектор значений факторов, определяющий точку в p+1 пространстве, в которой строим прогноз;

X – матрица, по которой было построено уравнение.

S – стандартное отклонение остаточной дисперсии или стандартная ошибка уравнения регрессии:

$$s = \sqrt{D_{\text{ост}}} \qquad (33)$$

Стандартная ошибка уравнения регрессии s может быть вычислена с помощью инструмента “Регрессия” надстройки «Пакет Анализа» MS Excel .

ПРИМЕР

Исследовать зависимость между стоимостью грузовой автомобильной перевозки Y (тыс. руб), весом груза X_1 (тонн) и расстоянием X_2 (тыс. км) по 20 транспортным компаниям (табл.3).

Таблица 3

№n/n	***Y***	***X_1***	***X_2***
1	51	35	2
2	16	16	1,1
3	74	18	2,55
4	7,5	2	1,7
5	33	14	2,4
6	26	33	1,55

Продолжение таблицы 3

№n/n	***Y***	X_1	X_2
7	*11,5*	20	0,6
8	*52*	25	2,3
9	*15,8*	13	1,4
10	*8*	2	2,1
11	*26*	21	1,3
12	*6*	11	0,35
13	*5,8*	3	1,65
14	*13,8*	3,5	2,9
15	*6,2*	2,8	0,75
16	*7,9*	17	0,6
17	*5,4*	3,4	0,9
18	*56*	24	2,5
19	*25,5*	9	2,2
20	*7,1*	4,5	0,95

Требуется построить и оценить линейную модель множественной регрессии по следующему плану:

1. Вычислить описательные статистики для отклика и всех факторов.

2. Оценить визуально, построив соответствующие облака рассеяния величины Y в зависимости от X_1 и X_2, целесообразность использования линейного уравнения регрессии.

3. Вычислить и проанализировать:

- линейные коэффициенты парной и частной межфакторной корреляции;
- линейные коэффициенты парной и частной корреляции между каждым фактором и откликом.

4. Написать уравнение множественной регрессии $Y = b_0 + b_1 X_1 + b_2 X_2$, оценить значимость его параметров, пояснить их экономический смысл. Коэффициенты уравнения вычислить двумя способами, используя:

- функцию ЛИНЕЙН();
- надстройку «Анализ Данных».

5. Написать уравнение множественной регрессии в стандартизованном масштабе $t_y = \beta_1 t_{x_1} + \beta_2 t_{x_2}$, пояснить экономический смысл его параметров.

6. Вычислить средние частные коэффициенты эластичности $\overline{Э}_{yx_1}$ и $\overline{Э}_{yx_2}$. Пояснить их экономический смысл.

6. Вычислить коэффициентом множественной детерминации $R^2_{yx_1x_2}$ двумя способами:

- по определению, по формуле (8);
- с использованием матрицы парных коэффициентов корреляции по формуле (9).

7. Полученный результат сравнить с результатом, полученным с помощью надстройки «Анализ Данных» $\Rightarrow$ «Регрессия».

8. С помощью *F*-критерия Фишера дать оценку надежности уравнения регрессии в целом и показателя тесноты связи R^2, используя результат, полученный с помощью надстройки «Анализ Данных»;

9. Оценить значимость коэффициентов множественной регрессии с помощью *t*-критерия Стьюдента, используя результаты работы надстройки «Анализ Данных» $\Rightarrow$ «Регрессия».

10. С помощью частных *F*-критериев Фишера оценить целесообразность включения в уравнение множественной регрессии фактора X_1 после фактора X_2 и фактора X_2 после фактора X_1.

11. Найти точечный и интервальный прогноз значений отклика при условии, что значение каждого фактора меньше максимального значения на 10% величины размаха исходных данных.

Пояснения по выполнению отдельных пунктов задания

Решение проведем с использованием электронных таблиц MS Excel.

К пункту 1

Исходные данные представлены на рис.1.*а*, и содержатся в интервале B3:D22, на рис.1.*б* приведены вычисленные средние значения, дисперсии и стандартные отклонения факторов и отклика.

	A	B	C	D	E	F	G
1	**Множественная регрессия.**						
2	N п.п.	y	x1	x2	yтеор	u-остатки	u^2
3	1	51	35	2	53,35668	-2,356683199	5,553956
4	2	16	16	1,1	17,798	-1,797996563	3,232792
21	19	25,5	9	2,2	26,32001	-0,820010329	0,672417
22	20	7,1	4,5	0,95	2,237743	4,862257249	23,64155
23	Суммы	454,5	277,2	31,8		1,59872E-14	
24	Средн.	22,725	13,86	1,59			
25	Ст.откл	19,8473896	10,04716	0,738004			

Рис.1.*а*. Исходные данные задачи в MS Excel

Описательные статистики для отклика и всех факторов X_1 и X_2, могут быть вычислены с помощью с помощью надстройки MS Excel «Пакет Анализа $\Rightarrow$ Описательные статистики» и представлены на рис. 1.*б*.

К пункту 2

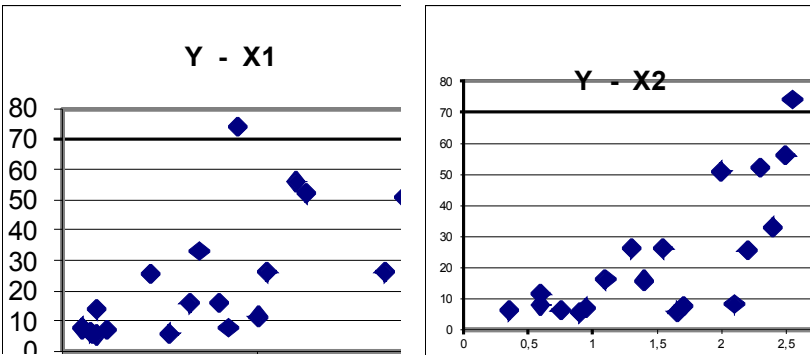
Вытянутость облака точек на диаграмме рассеяния (рис.2. *а*) вдоль наклонной прямой позволяет сделать предположение о том, что существует линейная связь между значениями переменных X_1 - весом груза и Y - стоимостью грузовой автомобильной перевозки.

Анализируя рис.2.*б*, можно заметить наличие прямой линейной связи между значениями переменных X_2 - расстоянием и Y - стоимостью грузовой автомобильной перевозки.

	I	J	L	N
3	**Описательная статистика**			
4				
5		*y*	*x1*	*x2*
6				
7	Среднее	22,725	13,86	1,59
8	Стандартная ошибка	4,5533	2,3050	0,1693
9	Медиана	14,80	13,50	1,60
10	Мода	26,00	2,00	0,60
11	Стандартное отклонение	20,3630	10,3082	0,7572
12	Дисперсия выборки	414,6514	106,2583	0,5733
13	Эксцесс	0,6974	-0,5669	-1,2001
14	Асимметричность	1,2794	0,5556	0,0035
15	Интервал	68,60	33,00	2,55
16	Минимум	5,40	2,00	0,35
17	Максимум	74,00	35,00	2,90
18	Сумма	454,50	277,20	31,80
19	Счет	20	20	20
20	**Прогноз**		31,70	2,65
21				
22				

=L17-L15*0,1

Рис.1 *б*. Описательная статистика для исходных данных задачи с помощью с помощью надстройки MS Excel «Пакет Анализа – Описательные статистики».



а) б)

Рис. 2 Облака рассеяния:

а) Y - X_1 и *б*) Y - X_2

К пункту 3

Значения **линейных коэффициентов парной корреляции** определяют тесноту попарно связанных переменных, использованных в данном уравнении множественной регрессии. **Линейные коэффициенты частной корреляции** оценивают тесноту связи значений двух переменных, исключая влияние всех других переменных, представленных в уравнении множественной регрессии. Матрицу парных коэффициентов корреляции переменных можно рассчитать, используя инструмент «Анализ данных» ⇒«Корреляция». Для этого:

1). В главном меню последовательно выберите пункты **Данные ⇒Анализ данных⇒Корреляция**. Щелкните по кнопке ***ОК***;

2). Заполните диалоговое окно ввода данных и параметров вывода (рис.3).

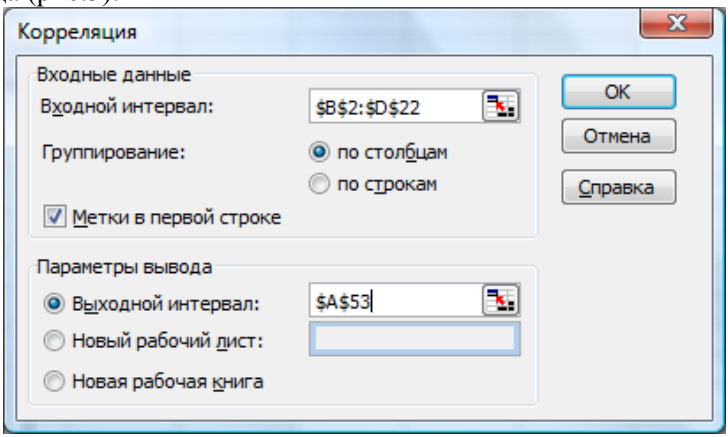


Рис.3 Диалоговое окно ввода данных и параметров вывода для вычисления коэффициентов парной корреляции

Значения **коэффициентов парной корреляции** указывают на заметную связь стоимости перевозок Y как с весом груза – X_1, так и расстоянием – X_2 (r_{yx1}=0,66 и r_{yx2}=0,63). В то же время межфакторная связь r_{x1x2} =0,12 довольно слабая, т.е. явной мультиколлинеарности нет. В связи с вышеизложенным, можно

сделать предварительный вывод, что нет оснований исключать факторы X_1 или X_2 из данной модели.

Коэффициенты частной корреляции дают более точную характеристику тесноты связи двух признаков, чем коэффициенты парной корреляции, так как очищают парную зависимость от взаимодействия данной пары признаков с другими признаками, представленными в модели (рис.4).

	A	B	C	D	E	F
51	**Надстройка-"Анализ данных"-" Корреляция"**					
52	**Значение коэффициентов парной корреляции**					
53		*y*	*x1*	*x2*		
54	y	1				
55	x1	0,6552333	1			
56	x2	0,6345813	0,124662	1		
57	**Коэффициенты частной корреляции**					
58	$r_{yx\,1\cdot x\,2} = \frac{r_{yx\,1} - r_{yx\,2}\cdot r_{x\,1\,x\,2}}{\sqrt{(1-r_{yx\,2}^{\,2})\cdot(1-r_{x\,1\,x\,2}^{\,2})}} =$					
59						**0,7513101**
60						
61	$r_{yx\,2\cdot x\,1} = \frac{r_{yx\,2} - r_{yx\,1}\cdot r_{x\,1\,x\,2}}{\sqrt{(1-r_{yx\,1}^{\,2})\cdot(1-r_{x\,1\,x\,2}^{\,2})}} =$					
62						**0,7376567**
63						
64	$r_{x\,1\,x\,2\cdot y} = \frac{r_{x\,1\,x\,2} - r_{yx\,2}\cdot r_{yx\,1}}{\sqrt{(1-r_{yx\,2}^{\,2})\cdot(1-r_{yx\,1}^{\,2})}} =$					
65						**-0,498662**
66						
67						

Рис.4 Результаты вычисления коэффициентов корреляции (интервал A54:D56) и коэффициентов частной корреляции

Вычислим коэффициенты частной корреляции по рекуррентным формулам:

$$r_{x1x2\cdot y} = \frac{r_{x1x2} - r_{yx2}\cdot r_{yx1}}{\sqrt{(1-r_{yx2}^{\,2})\cdot(1-r_{yx1}^{\,2})}} = -0.49$$

$$r_{yx2\cdot x1} = \frac{r_{yx2} - r_{yx1} \cdot r_{x1x2}}{\sqrt{(1 - r_{yx1}^{\ 2}) \cdot (1 - r_{x1x2}^{\ 2})}} = 0.73$$

$$r_{yx1\cdot x2} = \frac{r_{yx1} - r_{yx2} \cdot r_{x1x2}}{\sqrt{(1 - r_{yx2}^{\ 2}) \cdot (1 - r_{x1x2}^{\ 2})}} = 0.75$$

Наиболее тесно связаны Y и X_1 ($r_{yx1\ x2}$ =0,7513), связь Y и X_2 чуть слабее*:* $r_{yx2\ x1}$=0,7376.

Между факторная зависимость X_1 и X_2 не очень сильная $|r_{x1x2\ \ y}|$ = 0,4987 , что подтверждает отсутствие коллинеарности между факторами.

Если сравнить коэффициенты парной и частной корреляции, то можно увидеть, что из-за наличия между факторной зависимости они отличаются друг от друга:

r_{yx1}= 0,6552; $r_{yx1\ x2}$= 0,7513; r_{yx2}=0,6346; $r_{yx2\ x1}$=0, 7376.

Частные коэффициенты корреляции между Y и X_1, Y и X_2 свидетельствуют о более сильных взаимосвязях переменных, чем это показывают значения парных коэффициентов корреляции. Это произошло потому, что парный коэффициент корреляции r_{x1x2} занизил тесноту связи между Y и X_1 , Y и X_2

К пункту 4

Вычисление параметров линейного уравнения множественной регрессии.

Нахождение коэффициентов регрессии можно выполнить, используя функцию ЛИНЕЙН() (рис.5).

Нахождение коэффициентов регрессии можно провести с помощью инструмента «Анализ Данных» ⇒ «Регрессия» (рис.6).

Следует помнить, что в отличие от парной регрессии в диалоговом окне при заполнении параметра «входной интервал X» следует указать не один столбец, а все столбцы, содержащие значения факторных признаков.

Результат приведен на рис.7.

	A	B	C	D	E	F
99						
100	**Уравнение множественной регрессии.**					
101	Лин	15,10401	1,156057	-17,31332		
102		3,352977	0,24629	6,44711		
103		0,739853	10,98002	#Н/Д		
104		24,17386	17	#Н/Д		
105		5828,843	2049,535	#Н/Д		
106						
107		X2	X1	1		
108	Ура	Y=15,10*X2+1,16*X1-17,31				
109						

Рис.5. Результаты применения функции ЛИНЕЙН().

По результатам всех вычислений уравнение множественной регрессии имеет вида

$$Y = -17.3 + 1.16 \cdot X_1 + 15.10 \cdot X_2$$

Величины b_1 и b_2 указывают, что с увеличением значений X_1 и X_2 на единицу отклик увеличивается соответственно на 1,16 и на 15,10 тыс.руб.

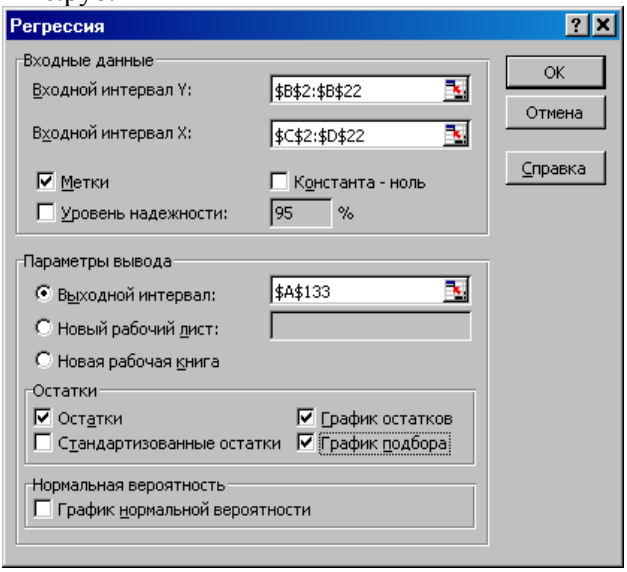


Рис..6 Диалоговое окно «Регрессия» инструмента «Анализ Данных»

	A	B	C	D	E	F	G
133	**ВЫВОД ИТОГОВ**						
134							
135	***Регрессионная статистика***						
136	**Множественный R**	**0,860147**					
137	**R-квадрат**	**0,739853**					
138	**Нормированный R-квадрат**	**0,709248**					
139	**Стандартная ошибка**	**10,98002**					
140	**Наблюдения**	**20**					
141							
142	**Дисперсионный анализ**						
143		***df***	***SS***	***MS***	***F***	***Значимость F***	
144	**Регрессия**	**2**	**5828,843**	**2914,421**	**24,17386**	**1,06993E-05**	
145	**Остаток**	**17**	**2049,535**	**120,5609**			
146	**Итого**	**19**	**7878,378**				
147							
148		***Коэффициенты***	***Стандартная ошибка***	***t-статистика***	***P-Значение***	***Нижние 95%***	***Верхние 95%***
149	**Y-пересечение**	**-17,31332**	**6,44711**	**-2,685439**	**0,015643**	**-30,91555429**	**-3,711089**
150	**x1**	**1,156057**	**0,24629**	**4,693892**	**0,000209**	**0,636430441**	**1,675683**
151	**x2**	**15,10401**	**3,352977**	**4,504657**	**0,000313**	**8,029837723**	**22,17818**
152							

Рис.7 Результаты применения инструмента «Анализ Данных» ⇒ «Регрессия»

К пункту 5

Для вычисления коэффициентов уравнения регрессии в стандартизованном масштабе используем формулы (6).

$$\beta_1 = b_1 \frac{\sigma_{X_1}}{\sigma_Y} = 1{,}16 \cdot \frac{10{,}05}{19{,}85} = 0{,}585,\ \beta_2 = b_2 \frac{\sigma_{2_1}}{\sigma_Y} = 15{,}10 \cdot \frac{0{,}74}{19{,}85} = 0{,}56$$

С учетом этого, уравнение регрессии в стандартном масштабе будет иметь вид:

$$t_Y = 0{,}58 t_{X_1} + 0{,}56 t_{X_2}$$

То есть, с ростом груза на одну сигму при неизменном расстоянии стоимость грузовых автомобильных перевозок увеличивается в среднем на 0,58 сигмы.

Поскольку значения коэффициентов отличаются друг от друга незначительно, то влияние на стоимость грузовых автомобильных обоих факторов приблизительно одинаково.

К пункту 6

Рассчитаем средние коэффициенты эластичности

$$\overline{Э}_{YX_1} = f'(\overline{X}_1)\frac{\overline{X}_1}{\overline{Y}} = b_1 \frac{\overline{X}_1}{b_0 + b_1\overline{X}_1 + b_2\overline{X}_2} = 1{,}16 \cdot 13{,}86 /$$

$$\left(-17{,}31 + 1{,}16 \cdot 13{,}86 + 15{,}10 \cdot 1{,}59\right) = 0{,}71$$

$$\overline{Э}_{YX_1} = f'(\overline{X}_2)\frac{\overline{X}_2}{\overline{Y}} = b_2 \frac{\overline{X}_2}{b_0 + b_1\overline{X}_1 + b_2\overline{X}_2} = 1{,}05 \text{ ё}$$

С увеличением среднего веса груза на 1% от его среднего уровня средняя стоимость перевозок возрастет на 0,71% от своего среднего уровня; при увеличении среднего расстояния перевозок на 1% - средняя стоимость доставки груза увеличится на 1,05%. Различия в силе влияния факторов на результат, полученные при сравнении уравнения регрессии в стандартизованном масштабе и коэффициентов эластичности, объясняются тем, что при вычислении коэффициентов эластичности учитывают поведение уравнения регрессии в окрестности средних значений.

К пункту 7

Величина коэффициента множественной детерминации R^2, рассчитанная по определению (по формуле (8) и с использованием коэффициентов уравнения множественной регрессии в стандартизованном масштабе (по формуле (9) оказалась одинаковой и равной 0.74. Расчеты приведены на рис.9.

Полученный результат совпадает с результатом, полученным с помощью надстройки «Анализ Данных» $\Rightarrow$«Регрессия», содержащимся в ячейке B136 («Множественный R») на рис.7, лист MS Excel в режиме отображения формул приводится на рис.8.

Поскольку коэффициент множественной детерминации оценивает долю вариации результата за счет представленных в уравнении факторов в общей вариации результата и $R^2_{yx_1x_2} = 0{,}7398$, то эта доля составляет 74 % и указывает на весьма высокую степень обусловленности вариации результата вариацией

факторов, иными словами – на весьма тесную связь факторов с результатом. О шкале Чеддока сила связи оценивается как высокая.

	A	B	C	D	E	F	G
181	**Вычисление коэффициента множественной детерминации**						
182							
183	**R^2=1-Sост/Sполн.**					**=**	**0,7399**
184							
185	**R^2=beta1*ryx1+beta2*ryx2**					**=**	**0,7399**
186							
187	**Вычисление исправленного коэффициента**						
188	**множественной детерминации**						
189							
190	**(R^2)adj=1-(1-R^2)*(n-1)/(n-p-1)**					**=**	**0,7092**

Рис.8 Вычисление коэффициента множественной детерминации в MS Excel в режиме отображения данных.

	A	F	G
181	**Вычисление коэффициента мн**		
182			
183	**R^2=1-Sост/Sполн.**	**=**	**=1-H23/E23**
184			
185	**R^2=beta1*ryx1+beta2*ryx2**	**=**	**=C156*C168+B156*D168**
186			
187	**Вычисление исправленного ко**		
188	**множественной детерминации**		
189			
190	**(R^2)adj=1-(1-R^2)*(n-1)/(n-p-1)**	**=**	**=1-(1-B137)*B146/B145**

Рис.9 Вычисление коэффициента множественной детерминации в MS Excel в режиме отображения формул.

К пункту 8

Оценку надежности уравнения регрессии в целом и показателя тесноты связи R^2 дает F- критерий Фишера.

По данным дисперсионного анализа, представленным в интервале ячеек A142:F146 на рис.10 , $F_{набл} = 24{,}16$.

Вероятность случайно получить такое значение F – критерия составляет $1{,}07*10^{-5}$, ячейка F144 («Значимость F»), что не превышает допустимый уровень значимости 5%. Следовательно, полученное значение не случайно, оно сформировалось под

влиянием существенных факторов, т.е. подтверждается статистическая значимость всего уравнения и показателя тесноты связи $R^2_{yx_1x_2}$.

К пункту 9

Оценку значимости коэффициентов множественной регрессии произведем с помощью ***t*** -критерия Стьюдента с использованием формулу (20).

Значения случайных ошибок параметров b_0, b_1 и b_2 с учетом округления:

m_{b0}=6,4471; m_{b1}=0,2463; m_{b2}=3,3530

Эти значения содержатся в диапазоне ячеек B102:D102 (рис.8), который является частью результата применения функции ЛИНЕЙН.

Эти значения используются для расчета t - критерия Стьюдента по формуле (20):

$$t_{b_0} = \frac{-17{,}3133}{6{,}4471} = -2{,}68;\ t_{b_1} = \frac{1{,}1560}{0{,}2463} = 4{,}69;\ t_{b_2} = 4{,}50.$$

Модули вычисленных величин следует сравнить с $t_{крит.}$, определяемым с заданным уровнем значимости и числом степеней свободы равным n-p-1.

Значение $t_{крит}$ равно 2.10. Поскольку модули значений наблюдаемых значений больше критического, гипотеза о равенстве нулю коэффициентов отвергается.

Это позволяет сделать вывод о существенности данного параметра, который формируется под воздействием неслучайных причин. Здесь статистически значимыми являются все коэффициенты b_0 , b_1 и b_2.

Процедура проверки значимости коэффициентов уравнения с помощью инструмента «Анализ Данных» $\Rightarrow$ «Регрессия» существенно проще, поскольку все промежуточные операции выполняются автоматически. На рис.7 в интервале B149:E151 приведены значения коэффициентов регрессии, стандартных

ошибок, t – статистики (t-наблюдаемые) и P – значения соответственно.

Для анализа значимости коэффициентов регрессии столбец «*P*-значение» в интервале E86:E88: если он меньше принятого нами уровня значимости (в настоящей работе уровень значимости принят равным 0,05), делают вывод о неслучайной природе данного значения коэффициента, т.е. о том, что он статистически значим и надежен. В противоположном случае принимается гипотеза о случайной природе значения этого коэффициента уравнения. Здесь все P<0,05, что позволяет подтвердить сделанный ранее вывод о статистической значимости всех параметров регрессии

Очень важно помнить, что все найденные коэффициенты b_0, b_1 и b_2.являются точечными оценками коэффициентов регрессии, при этом интервальными оценками с надежностью 95% являются интервалы, границы которых указаны в таблице 4.

Таблица 4

Коэффициент	Интервал	
	Левая граница	Правая граница
b_0	-30,92	-3,71
b_1	0,636	1,676
b_2	8,030	22,178

Знание этих интервалов позволяет получить важные выводы о том, что с увеличением веса груза на одну тонну (при неизменном значении расстояния) с вероятностью 95% стоимость поездки в среднем возрастет на величину от 0,636 до 1,676 тыс.руб. Увеличение расстояния на одну тыс. км (при неизменном значении веса груза) с вероятностью 95% увеличивает в среднем стоимость поездки на величину от 8,030 до 22,178 тыс.руб.

К пункту 10

Для оценки целесообразности включения в модель фактора X_1 после фактора X_2 и фактора X_2 после фактора X_1 вычислим значения частных F-критериев Фишера:

$$F_{\text{частн } x_2} = \frac{R^2_{yx1x2} - R^2_{yx1}}{1 - R^2_{yx1x2}} \cdot \frac{n-p-1}{1} = \frac{0.7398 - 0.6552^2}{1 - 0.7398} \cdot \frac{20-2-1}{1} = 20.29$$

$$R^2_{yx1=} r^2_{yx1}$$

$$F_{\text{частн } x_1} = \frac{R^2_{yx1x2} - R^2_{yx2}}{1 - R^2_{yx1x2}} \cdot \frac{n-p-1}{1} = \frac{0.7398 - 0.6345^2}{1 - 0.7398} \cdot \frac{20-2-1}{1} = 22.03$$

$$R^2_{yx2=} r^2_{yx2}$$

Частный F-критерий – $F_{\text{частн } x_2}$ показывает статистическую зависимость включения фактора X_2 в модель после того, как в нее включен фактор X_1 . $F_{\text{частн } x_2}$ = 20,29. Найдем $F_{крит}$=4,45 при принятом уровне значимости a =0,05 (5%) (число степеней свободы числителя и знаменателя равны 1 и 17 соответственно).

$F_{\text{частн } x_2}$ = 20,29 > $F_{\text{крит}}$ =4,45. Следовательно, включение в модель фактора X_2 – расстояния, после того, как уравнение включен фактор X_1 – вес груза, статистически целесообразно: прирост факторной дисперсии за счет дополнительного признака X_2 оказывается значительным, существенным; фактор X_2 следует включать в уравнение после фактора X_1 .

Поменяем первоначальный порядок включения факторов в модель и рассмотрим вариант включения X_1 после X_2 . Для этого вычислим $F_{\text{частн } x_1}$, оно равно 22,03 при том же уровне значимости a =0,05 (5%). $F_{\text{крит}}$ =4,45 и $F_{\text{частн } x_1} > F_{\text{крит}}$. Следовательно, значение частного F -критерия для дополнительно включенного фактора X_1 не случайно, является статистически значимым, надежным, достоверным: прирост факторной дисперсии за счет дополнительного фактора X_1 является существенным.

Фактор X_1 должен присутствовать в уравнении, в том числе в варианте, когда он дополнительно включается после фактора X_2 .

К пункту 11

Вычислим значения каждого фактора $X_1^{\text{прогн.}}$ (тонн) - вес груза и $X_2^{\text{прогн.}}$ (тыс. км) - расстояние, в которых будем строить прогноз. В качестве прогнозных значений возьмем величину равную $x_{\max} - 0.1 \cdot (x_{\max} - x_{\min})$, где $x_{\max}$ и $x_{\min}$ максимальное и минимальное значения факторов в таблице исходных данных. Вычислим прогнозные значения каждого фактора с учетом того, что максимальные значения и размахов $(x_{\max} - x_{\min})$ уже были вычислены и приведены на рис.1.*б*. $X_1^{\text{прогн.}} = 31{,}7$; $X_2^{\text{прогн.}} = 2{,}645$.

Все вычисления приведены на рис.10-16. Заметим, что перед вычислением все необходимые данные были скопированы на новую страницу MS Excel. Для расчета точечного прогноза $\hat{Y}$ подставим полученные результаты вычислений в уравнение множественной регрессии

$$\hat{Y} = -17.3 + 1.16 \cdot X_1^{прогн.} + 15.10 \cdot X_2^{прогн.} =$$

$$= -17.3 + 1.16 \cdot 31.7 + 15.10 \cdot 2.645 = 59.28$$

Полученное прогнозное значение будет серединой интервального прогноза. Для того, чтобы вычислить ширину доверительного интервала необходимо вычислить выражение (30). Вектор $X0$ расположен в интервале C2:C4 (рис.10) и содержит прогнозные значения факторов, при этом первый элемент всегда равен единице.

	A	B	C
1		Прогнозирование	
2			1
3		X0=	31,7
4			2,645
5			
6	X		
7	x0	x1	x2
8	1	35,00	2,00
9	1	16,00	1,10
26	1	9,00	2,20
27	1	4,50	0,95

Рис.10 Расчет интервального прогноза. Режим отображения данных. (начало)

Матричные операции для вычисления $\sqrt{(X0)^T\left(X^TX\right)X0}$ проведены на рис. 12.

	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W	X	Y	Z
1																							
2	X0__transp	1	31,7	2,645																			
3																							
4																							
5			X_transp																				
6			x0	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
7			x1	35	16	#	2	#	#	#	#	2	#	#	3	4	3	#	3	#	9	4,5	
8			x2	2	1,1	3	2	2	2	1	2	1	2	1	0	2	3	1	1	1	3	2,2	0,95

Рис.11 Расчет интервального прогноза. Режим отображения данных. Транспонирование матриц (продолжение)

	F	G	H	I	J	K	L
14	X_transp*X				(X_transp*X)^(-1)		
15	20,0	277,2	31,8		0,3	0,0	-0,1
16	277,2	5860,9	459,2		0,0	0,0	0,0
17	31,8	459,2	61,5		-0,1	0,0	0,1
18							
19							
20							
21							
22	X0_transp*(X_transp*X)^(-1)						
23	-0,194	0,0081	0,083				
24							
25	X0_transp*(X_transp*X)^(-1)*X0						
26	0,282						
27	КОРЕНЬ(X0_*(X_transp*X)^(-1)*X0)						
28	0,531						

Рис.12 Расчет интервального прогноза. Режим отображения данных.

Вычисление $\sqrt{(X0)^T\left(X^TX\right)X0}$ (продолжение)

В ячейку G34 записано значение s – стандартное отклонение остаточной дисперсии или стандартная ошибка уравнения регрессии, которая была получена с помощью инструмента «Анализ Данных» ⇒ «Регрессия» содержится в ячейке G139 (рис.7).

Искомая величина ε – половина ширины доверительного интервала вычислена по формуле (30) и содержится в ячейке G36 рис.13.

	F	G	H	I	J	K
27	КОРЕНЬ(X0_*(X_transp*X)^(-1)*X0)					
28	0,531					
29	Gamma	0,95				
30	n=	20				
31	p=	2				
32	t_крит=	2,1098				
33						
34	s=	10,98				
35						
36	eps=	12,297				
37						
38						
39		Коэффициент		Точечный прогноз		
40	Y-пере	-17,313		Y_прогн=	59,3	
41	X1	1,1561				
42	X2	15,104		Доверительный интервал		
43				Y_нижн		Y_верхн
44				46,987		71,581

Рис.13 Расчет точечного и интервального прогноза. Окончание. Режим отображения данных

Границы доверительного интервала вычислены с использованием формулы (29) и помещены в ячейки I44 и K44.

Все вычисления в режиме отображения формул приведены на рис.14 -16.

	F	G	H	I	J	K	L	M
14	X_transp*X						(X_transp*X)^(-1)	
15	=МУМНОЖ(A8:C27;G6:Z8)	=МУМНОЖ(A	=МУМНОЖ(A8:C27;G6:Z8)				=МОБР(F15:H17)	=M(
16	=МУМНОЖ(A8:C27;G6:Z8)	=МУМНОЖ(A	=МУМНОЖ(A8:C27;G6:Z8)				=МОБР(F15:H17)	=M(
17	=МУМНОЖ(A8:C27;G6:Z8)	=МУМНОЖ(A	=МУМНОЖ(A8:C27;G6:Z8)				=МОБР(F15:H17)	=M(
22	X0_transp*(X_transp*X)^(-1)							
23	=МУМНОЖ(E2:G2;L15:N17)	=МУМНОЖ(E	=МУМНОЖ(E2:G2;L15:N17)					
24								
25	X0_transp*(X_transp*X)^(-1)*X0							
26	=МУМНОЖ(F23:H23;C2:C4)							
27	КОРЕНЬ(X0_*(X_transp*X)^(-1)*X0)							
28	=КОРЕНЬ(F26)							

Рис.14 Расчет интервального прогноза. Режим отображения формул. Операции с матрицами и вычисление $\sqrt{(X0)^T\left(X^TX\right)X0}$ (продолжение)

	F	G
28	=КОРЕНЬ(F26)	
29	Gamma=	0,95
30	n=	20
31	p=	2
32	t_крит=	=СТЬЮДРАСПОБР(1-G29;G30-G31-1)
33		
34	s=	10,98002099
35		
36	eps=	=F28*G32*G34
37		
38		
39		Коэффициент
40	Y-пересечение	-17,31332187
41	X1	1,156056725
42	X2	15,10400985

Рис.15 Расчет интервального прогноза. Режим отображения формул. Вычисление ширины интервала (продолжение)

	I	J	K	L
39	Точечный прогноз			
40	Y_прогн=	=МУМНОЖ(E2:G2;G40:G42)		
41				
42	Доверительный интервал			
43	Y_нижн		Y_верхн	
44	=J40-G36		=J40+G36	

Рис.16. Расчет точечного и интервального прогноза. Окончание. Режим отображения формул.

Интервальной оценкой прогноза с указанными значениями факторов является доверительный интервал с надежностью 95 % [46,99 ; 71,59] тыс.руб.

Так, стоимость перевозки груза весом 13,86 тонн на расстояние 1,59 тыс. км с вероятностью 95 % будет лежать в пределах [46,99 ; 71,59].

Общий вывод состоит в том, что множественная линейная модель

$$\hat{Y} = -17{,}31 + 1{,}16 \cdot X_1 + 15{,}10 \cdot X_2$$

с факторами X_1 и X_2 имеет коэффициент детерминированности $R^2_{yx_1x_2}$ =0,73. Она содержит информативные факторы X_1 и X_2.

Уравнение парной регрессии является простым, хорошо детерминированным, пригодным для анализа и для прогноза

.

ЗАДАНИЕ

Для ряда регионов представлена информация об объёмах Y **(у.е.)** продаж фирмы «Галактика» и ее затратах на рекламу в этих регионах – X_1, а также индекс потребительских доходов в этих регионах – X_2. Построить и оценить линейную модель множественной регрессии по плану, приведенному в примере, изложенном выше.

Исходные данные взять из файла «*LabRab_6.xls*».

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Эконометрика. Парная линейная регрессия. Методические указания к лабораторным работам для студентов направлений подготовки бакалавриата 21.03.02 и 38.03.01 / Сост.: *В.В. Беляев, Т.Р. Косовцева.* СПб, 2016. - 50 с.
2. *Магнус Я.Р.* Эконометрика. Начальный курс. Учебник для вузов. - / Я.Р. Магнус, П.К. Катышев, А.А. Пересецкий; М., «Дело» 2000. - 400 с.
3. Эконометрика./ Учебник для бакалавров. Под ред. Елисеевой И.И., М., «Проспект», 2014. - 288 с..
4. Практикум по эконометрике./ Под редакцией Елисеевой И.И., М., «Финансы и статистика», 2004. - 192 с.
5. *Тихомиров Н.* Эконометрика. Учебник / Н. Тихомиров, Е.М. Дорохина: М.: «Экзамен», 2006 . - 512 с.
6. *Кремер Н. Ш.* Эконометрика. Учебник для вузов, / Н.Ш. Кремер, Б.А. Путко М.: М.: Юнити, 2005. - 311 с.
7. *Арженовский С.В*. Эконометрика: учебное пособие/ С.В. Арженовский, О.Н. Федосова. Рост.гос.экон.университет – Ростов н/Д., 2002. - 102 с.

СОДЕРЖАНИЕ

ЭКОНОМИКО-МАТЕМАТИЧЕСКИЕ МЕТОДЫ И МОДЕЛИРОВАНИЕ

МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

Методические указания к лабораторным работам для студентов бакалавриата направления 21.03.02

Сост.: *В.В. Беляев, Т.Р. Косовцева*

Печатается с оригинал-макета, подготовленного кафедрой информатики и компьютерных технологий

Ответственный за выпуск *Т.Р. Косовцева*

Лицензия ИД № 06517 от 09.01.2002

Подписано к печати 22.01.2020. Формат 60×84/16.
Усл. печ. л. 2,4. Усл.кр.-отт. 2,4. Уч.-изд.л. 1,8. Тираж 75 экз. Заказ 18. С 1.

Санкт-Петербургский горный университет
РИЦ Санкт-Петербургского горного университета
Адрес университета и РИЦ: 199106 Санкт-Петербург, 21-я линия, 2