Белорусский национальный технический университет

Приборостроительный факультет

Кафедра «Инженерная математика» СОГЛАСОВАНО Заведующий кафедрой «Инженерная математика» _____М.А. Князев 24 апреля 2024 г.	СОГЛАСОВАНО Декан приборостроительного факультета _____А.И.Свистун 29 апреля 2024 г.

Электронный учебно-методический комплекс

по учебной дисциплине

«ПРИКЛАДНАЯ МАТЕМАТИКА»

для студентов специальности
6-05-0716-01 «Метрология, стандартизация и контроль качества»

Составители:
старший преподаватель кафедры «Инженерная математика» Прихач Наталия Константиновна, доцент кафедры «Инженерная математика» Прусова Ирина Васильевна, доцент кафедры «Инженерная математика» Романчак Василий Михайлович

Рассмотрено и утверждено
на заседании Совета ПСФ
Протокол № 8 29 апреля 2024 года

Минск БНТУ 2024

Перечень материалов

Электронный учебно-методический комплекс (ЭУМК) по учебной дисциплине «Прикладная математика» состоит из следующих разделов:

I. Теоретический раздел:
– учебные материалы (конспект лекций).

II. Практический раздел:
– лабораторные работы по курсу.

III. Контроль знаний
– перечень вопросов, выносимых на зачет (экзамен),
– проверочные тесты по темам курса «Прикладная математика» с вариантами ответов.

IV. Вспомогательный раздел:
– учебная программа для учреждения высшего образования.

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА
Цели создания ЭУМК:
Целью ЭУМК по дисциплине «Прикладная математика» является формирование у студентов комплекса знаний по изучаемой учебной дисциплине, соответствующих академическим, социально-личностным и профессиональным компетенциям специалиста в рамках образовательного стандарта для специальности приборостроительного факультета 6-05-0716-01.
Особенностями структурирования и подачи учебного материала являются изучение следующих теоретических материалов:
- базовые теоретические сведения по дисциплинам «Математика» и «Информатика».
- практические навыки по теме «Случайные величины и законы их распределения»;
Практическая часть состоит из обучающих материалов для выполнения необходимых разделов курсовой работы, задач для самостоятельного решения.
Раздел контроля знаний включает проверочные тесты с вариантами ответов, вопросы к зачету (экзамену).
Вспомогательный раздел содержит учебную программу по дисциплине «Информатика».
Рекомендации по организации работы с ЭУМК: Материалы данного электронного учебно-методического комплекса можно использовать в качестве методической поддержки при проведении лабораторных занятий для закрепления и углубления знаний, для самостоятельной работы и подготовки студентов к зачету (экзамену)**.**

Содержание

I ТЕОРЕТИЧЕСКИЙ РАЗДЕЛ
1.1. КОНСПЕКТ ЛЕКЦИЙ

РАЗДЕЛ I Случайные величины
ТЕМА 1.1 Основные понятия теории вероятностей. Случайные величины и их числовые характеристики
1.1.1 Случайные события и вероятность

События, происходящие в окружающем мире можно разделить на три вида: достоверные, невозможные и случайные.

Достоверным называется событие, которое обязательно произойдет, которое обязательно произойдет, если будет осуществлена определенная совокупность условий.

Событие называется *невозможным*, если оно никогда не произойдет при совокупности данных условий.

Событие называется *случайным*, если в результате наблюдения или испытания оно может произойти или не произойти.

Соблюдение совокупности условий при проведении эксперимента называется *испытанием*. Событие является результатом (или *исходом*) испытания.

События называются *несовместными*, если появление одного из них исключает появление других событий в одном и том же испытании.

Несколько событий образуют *полную группу*, если в результате испытания появится хотя бы одно из них. При этом, если события, образующие полную группу, попарно несовместны, то в результате испытания появится одно и только одно из этих событий.

Противоположным называют два единственно возможных события, образующих полную группу. Если одно из двух противоположных событий обозначено через A, то другое принято обозначать $\overline{A}$.

События называются *равновозможными*, если есть основание считать, что ни одно из них не является более возможным, чем другое.

Основной характеристикой случайного события является его вероятность. *Вероятностью события A* называется отношение числа *m*, благоприятствующих этому событию исходов опыта, к общему числу *n* всех несовместных единственно возможных и равновозможных исходов:

$$P(A) = \frac{m}{n} \qquad (1.1)$$

Из определения вероятности вытекают следующие свойства:

1) вероятность достоверного события равна 1;
2) вероятность невозможного события равна 0;

3) вероятность случайного события есть положительное число, заключенное между нулем и единицей.

Суммой $A+B$ *событий* A и B называют событие, состоящее в появлении события A, или события B, или обоих этих событий.

Теорема 1.1. Вероятность появления одного из двух несовместных событий, безразлично какого, равна сумме вероятностей этих событий:

$$P(A+B)=P(A)+P(B).$$

Теорема 1.2. Сумма вероятностей несовместных событий $A_1, A_2, \ldots, A_n$, образующих полную группу, равна единице:

$$P(A_1)+P(A_2)+\ldots+P(A_n)=1.$$

Теорема 1.3. Сумма вероятностей противоположных событий равна единице:

$$P(A)+P(\overline{A})=1.$$

Произведением двух событий A и B называют событие $A\cdot B$, состоящее в их совместном появлении.

Если при вычислении вероятности события никаких других ограничений, кроме необходимого комплекса условий испытания, не налагается, то такая вероятность называется безусловной. Если же налагаются другие дополнительные условия, то вероятность события называется *условной*.

Условной вероятностью $P_A(B)$ называют вероятность события B, вычисленная в предположении, что событие A уже наступило.

Теорема 1.4. Вероятность произведения двух событий определяется формулой $P(A\cdot B)=P(A)\cdot P_A(B)$.

1.1.2 Способы описания и характеристики случайных величин

Случайная величина (СВ) – величина, которая в результате опыта может принять то или иное значение, причём неизвестно заранее, какое именно. Обычно случайные величины обозначаются заглавными буквами латинского алфавита *X, Y, Z* и т.д., а конкретные их значения – маленькими буквами, например $x_1, x_2, x_3, \text{К}$, $y_1, y_2, y_3, \ldots, z_1, z_2, z_3$.

Случайные величины могут быть *дискретные* (например, количество студентов на лекции) и *непрерывные* (например, температура окружающей среды или давление).

На практике больше приходится иметь дело со случайными величинами, чем со случайными событиями. Для того чтобы все знать о случайной величине, нужно знать закон распределения вероятностей.

Закон распределения вероятностей случайной величины (*или просто закон распределения*) – это соотношение, устанавливающее связь между возможными значениями случайной величины и соответствующими им вероятностями. Этот закон может быть задан в виде таблицы, формулы или графически. В случае табличного задания закона распределения первая строка таблицы указывает возможные значения случайной величины, а вторая – их вероятности, причём $\sum_{i=1}^{n} p_i = 1$:

Таблица 1.1.Закон распределения дискретной случайной величины

X	x_1	x_2	…	x_n
p	p_1	p_2	…	p_n

В случае графического задания дискретной случайной величины по оси *OX* откладывают значения этой величины, а по оси *OY* – соответствующие вероятности. Ломаная линия, соединяющая точки (x_i, y_i), называется *многоугольником распределения*.

Закон распределения дискретной случайной величины может быть задан также в виде формулы, отражающей зависимость вероятности от значения случайной величины: $p_i = P(x_i)$.

Функцией распределения случайной величины X называется функция $F(x)$, определяющая вероятность того, что случайная величина X в результате испытания примет значения, меньшее x: $F(x) = P(X < x)$.

Иногда вместо термина «функция распределения» используют термин «интегральная функция».

График функции распределения дискретной случайной величины имеет ступенчатый вид.

С помощью функции распределения можно описать не только дискретную, но и непрерывную случайную величину.

Непрерывную случайную величину можно также задать с помощью функции плотности распределения (или дифференциальной функции распределения).

Плотностью распределения вероятностей непрерывной случайной величины X называется первая производная от функции распределения:

Функция распределения является первообразной для плотности распределения, поэтому: $P(X < x) = F(x) = \int_{-\infty}^{x} f(z)dz$.

Таким образом, для описания дискретной случайной величины используется закон распределения и функция распределения, а для описания непрерывной случайной величины – функция распределения и функция плотности распределения.

Наряду с этим оба типа случайных величин характеризуются с помощью некоторых чисел, которые описывают случайную величину в целом. Такие числа называются *числовыми характеристиками случайной величины*. К ним относятся *математическое ожидание*, дисперсия, среднеквадратическое отклонение и др.

Математическое ожидание – это одно из важнейших понятий в теории вероятностей и математической статистике, характеризующее распределение значений или *вероятностей* случайной величины. Широко применяется при проведении технического анализа, исследовании числовых рядов, изучении непрерывных и продолжительных процессов. Имеет важное значение при оценке рисков, прогнозировании ценовых показателей при торговле на финансовых рынках, используется при разработке стратегий и методов игровой тактики в теории азартных игр.

Математической ожидание приближённо равно среднему значению случайной величины. Оно может быть рассчитано по следующим формулам:

$M(X) = \sum_{i=1}^{n} x_i p_i$ – для дискретной случайной величины с конечным множеством значений;

$M(X) = \int_{-\infty}^{\infty} x \cdot f(x)dx$ – для непрерывной случайной величины.

$M_0(X)$ дискретной случайной величины называется наиболее вероятное значение. Для непрерывной случайной величины мода – это точка максимума ее плотности вероятности.

Медианой $Me(X)$произвольной случайной величины называется такое ее значение, относительно которого равновероятно получение большего или меньшего значения.

На практике часто требуется оценить рассеяние возможных значений случайной величины вокруг ее среднего значения. Например, в артиллерии важно знать, насколько кучно лягут снаряды вблизи цели, которая должна быть поражена.

Одной из важных характеристик рассеяния является дисперсия.

Дисперсия характеризует *степень рассеяния* значений случайной величины вокруг ее математического ожидания. Она равна математическому ожиданию квадрата отклонения случайной величины от ее математического ожидания: $D(X) = M\left(X - M(X)\right)^2 = M\left(X^2\right) - \left[M(X)\right]^2$.

Если все значения случайной величины тесно сконцентрированы около её математического ожидания и большие отклонения от математического отклонения маловероятны, то такая случайная величина имеет малую дисперсию.

Если значения случайной величины рассеяны и велика вероятность больших отклонений от математического ожидания, то такая случайная величина имеет большую дисперсию.

Основным недостатком дисперсии в качестве характеристики разброса является то, что если случайная величина выражена в некоторых единицах измерения, то дисперсия имеет наименование, выраженное в квадратных единицах. Для удобства представления СВ через свои характеристики вводят понятие среднего квадратического отклонения $\sigma(X)$ (СКО), равное положительному арифметическому корню из дисперсии:

$$\sigma(X) = \sqrt{D(X)}\,.$$

Асимметрией и *эксцессом* дискретной случайной величины X называются соответственно величины:

$$As(X) = \frac{1}{\sigma_3}\sum_i p_i\left(x_i - M(X)\right)^3; \qquad Ex(X) = \frac{1}{\sigma^4}\sum_i p_i\left(x_i - M(X)\right)^4 - 3\,.$$

Знаки $As(X)$ и $Ex(X)$ указывают на отклонения графика закона распределения X от нормального распределения, для которого $As(X) = 0$; $Ex(X) = 0$ (см. рис. 1.1).

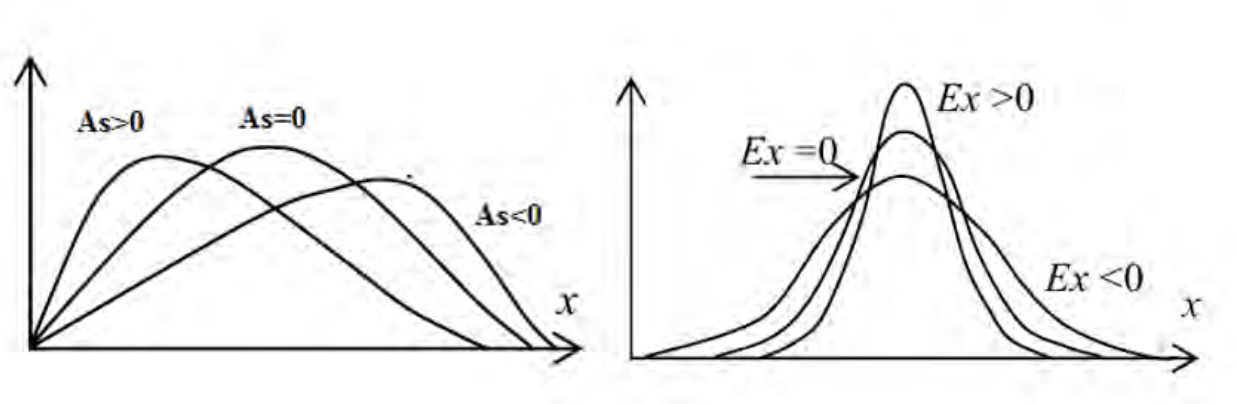


Рис. 1.1 – Графики $f(x)$ при разных знаках асимметрии и эксцесса

1.1.3 Основные виды распределений случайных величин

1. *Равномерное распр*еделение. Непрерывная случайная величина X называется *равномерно распределенной на интервале* $[a,b]$, если ее плотность вероятности равна константе $\frac{1}{b-a}$ на этом интервале и нулю вне его:

$$f(x)=\begin{cases}0, & x<a;\\ \dfrac{1}{b-a}, & a\le x\le b;\\ 0, & x>b.\end{cases}$$

Соответствующая функция распределения вероятностей имеет следующий вид:

$$F(x)=\begin{cases}0, & x<a;\\ \dfrac{x-a}{b-a}, & a\le x\le b;\\ 1, & x>b.\end{cases}$$

На рис. 1.2 изображены графики функции распределения и плотности распределения вероятности равномерно распределенной непрерывной случайной величины.

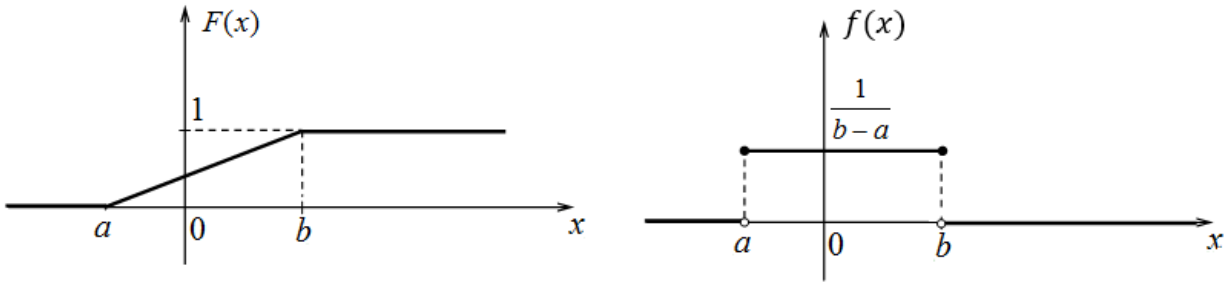


Рис. 1.2 – Графики функции распределения и плотности распределения вероятности равномерно распределенной СВ

2. *Экспоненциальное (показательное) распределение*. Случайная величина X распределена по экспоненциальному закону, если ее плотность распределения вероятностей определяется формулой:

$$f(x)=\begin{cases}0, & x<0;\\ \lambda\cdot e^{-\lambda x}, & x\ge 0,\end{cases}$$

где $\lambda=\text{const},\ \lambda>0$ – параметр распределения.

Функция распределения показательного распределения имеет вид:

$$F(x)=\begin{cases}0, & x<0;\\ 1-e^{-\lambda x}, & x\ge 0.\end{cases}$$

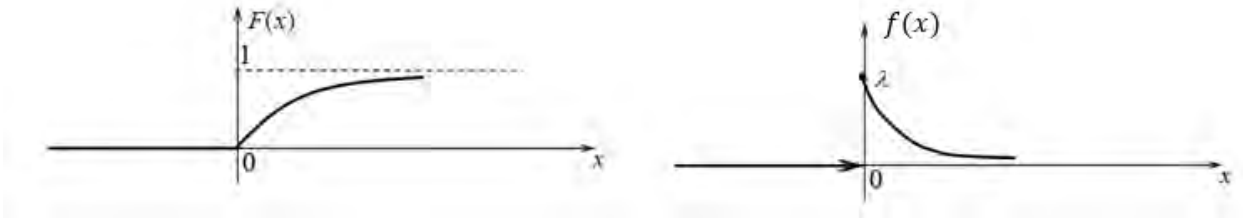


Рис. 1.3 – Графики функции распределения и плотности распределения вероятности СВ, распределенной по показательному закону

3. *Нормальное распределение.*

Нормальным распределением (или з*аконом Гаусса*) называется распределение непрерывной случайной величины X, *плотность* которой определяется по формуле:

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

где a и σ – параметры распределения.

Указанная функция является функцией Гаусса специального вида, поэтому ее график называют *нормальной* или *гауссовой* кривой. Эта кривая изображена на рис. 1.4.

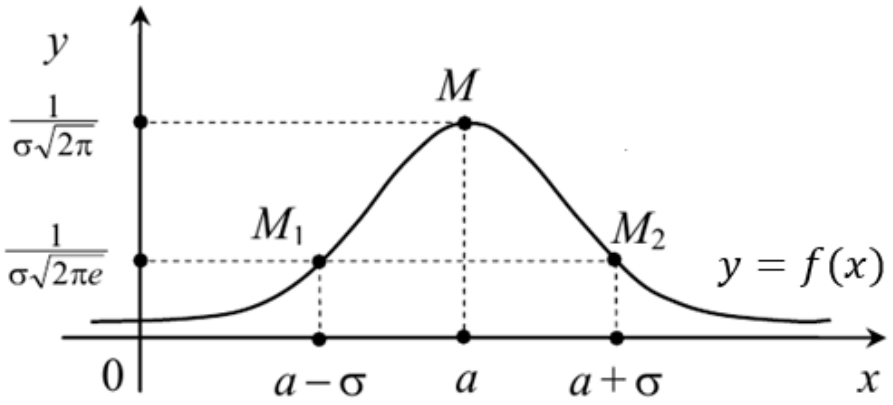


Рис. 1.4 – Кривая Гаусса

Основные числовые характеристики: математическое ожидание $M(X)=a$; дисперсия $D(X)=\sigma^2$, асимметрия $As=0$; эксцесс $Ex=0$.

Так как нормальное распределение симметрично относительно своего математического ожидания, то мода, медиана и математическое ожидание нормально распределенной случайной величины равны между собой.

Функция распределения $F(x)$ случайной величины X вычисляется по формуле: $F(x)=\frac{1}{2}+\Phi\left(\frac{x-a}{\sigma}\right)$, где $\Phi(t)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{t}e^{-\frac{z^2}{2}}dz$.

График функции $F(x)$ изображен на рис. 1.5 .

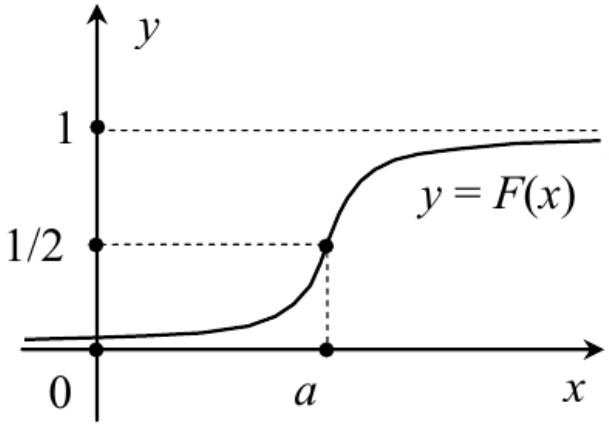


Рис. 1.5 – График функции $F(x)$

4. *Распределения, связанные с нормальным*.

1. *Распределение* χ^2.

Пусть $X_i; i=\overline{1,n}$ – независимые нормально распределённые случайные величины, математическое ожидание каждой равно 0, СКО – 1. Тогда величина $\chi^2=\sum_{i=1}^{n}X_i^2$ распределена по закону «хи-квадрат» со $k=n$ степенями свободы (число степеней свободы есть параметр распределения). *Число степеней свободы случайной величины* χ^2 определяется числом независимых случайных величин в сумме $\sum_{i=1}^{n}\chi_i^2$.

Функция плотности распределения χ^2 имеет вид:

$$f(x)=\begin{cases}0 \quad if \quad x\le 0\\ \dfrac{1}{2^{k/2}\Gamma(k/2)}e^{-x/2}x^{\frac{k}{2}-1} \text{ если } x>0,\end{cases} \text{ где } \Gamma(x)=\int_0^{\infty}t^{x-1}e^{-t}dt$$ – гамма-функция Эйлера. В частности, $\Gamma(n+1)=n!$ – где n – число степеней свободы.

Для χ^2- распределения составлены таблицы.

χ^2- распределение всегда неотрицательно. Зависит от числа степеней свободы. Математическое ожидание равно n – количеству переменных, а стандартное отклонение равно $2n$. С увеличением числа степеней свободы χ^2- распределение стремится к нормальному.

2. *Распределение Стьюдента***.** Пусть случайная величина T задана равенством $T = \dfrac{Z\sqrt{k}}{V}$, где Z – нормально распределённая случайная величина с нулевым математическим ожиданием и СКО, равным 1, а V – независимая от Z случайная величина, имеющая распределение χ^2 с k степенями свободы. Тогда T подчиняется распределению Стьюдента с k степенями свободы (параметр формы – положительное целое число), которое часто называют *t-распределением***.** Плотность вероятности имеет вид:

$$f(t,k) = \frac{\Gamma\left(\dfrac{k+1}{2}\right)}{\sqrt{k\pi}\cdot\Gamma\left(\dfrac{k}{2}\right)}\cdot\left(1+\frac{t^2}{k}\right)^{-\frac{k+1}{2}},\ -\infty < t < \infty.$$

Кривая плотности распределения Стьюдента симметрична относительно оси ординат. Она напоминает кривую плотности стандартного нормального распределения, но убывает несколько медленнее.

Математическое ожидание и дисперсия равны 0 и $\dfrac{k}{k-2}$ соответственно, при $k > 2$.

По сравнению с нормальным распределение Стьюдента более пологое, оно имеет большую дисперсию.

При увеличении числа степеней свободы кривая плотности приближается к кривой стандартного нормального распределения.

3. *Распределение Фишера (Фишера-Снедекора)*.

Пусть U и V – независимые случайные величины, распределённые по закону χ^2 со степенями свободы k_1 и k_2, тогда величина $F = \dfrac{U/k_1}{V/k_2}$ имеет распределение, которое называют *F-распределение* или *распределение Фишера-Снедекора* со степенями свободы k_1 и k_2.

Функция плотности распределения имеет вид:

$$f(x) = \begin{cases} 0, & x \le 0 \\ C_0 \cdot \dfrac{x^{(k_1-2)/2}}{(k_2+k_1x)^{(k_1+k_2)/2}}, & x > 0 \end{cases}, \text{ где } C_0 = \frac{\Gamma\left(\dfrac{k_1+k_2}{2}\right)\cdot k_1^2\cdot k_2^2}{\Gamma\left(\dfrac{k_1}{2}\right)\cdot\Gamma\left(\dfrac{k_2}{2}\right)}.$$

Основные числовые характеристики *F-распределения.*

$$m_F = \frac{k_2}{k_2 - 2},\ k_2 > 2 \quad Mo = \frac{k_2(k_1 - 2)}{k_1(k_2 + 2)},\ k_1 \geq 2 \quad D_F = \frac{2(k_1 + k_2 - 2)k_2^2}{k_1(k_2 - 2)^2(k_2 - 4)}$$

РАЗДЕЛ II Выборка и ее анализ

ТЕМА 2.1 Статистические оценки параметров распределения

2.1.1 Вариационные ряды и их графическое изображение

Генеральной совокупностью называется совокупность объектов произвольной природы, обладающих признаками, доступными для наблюдения и количественного измерения.

Объекты, входящие в генеральную совокупность, называются ее *элементами*, а их общее число N – ее *объемом*.

Однако, получение экспериментальных данных достаточно трудоемкий, дорогой процесс, а в некоторых случаях и просто невозможный. Поэтому из всей генеральной совокупности приходится выбирать только определенную часть объектов, которую называют *выборочной совокупностью* или *выборкой объема n.*

Предположим, что над случайной величиной X производится ряд независимых опытов (наблюдений). В каждом из этих опытов случайная величина X принимает определенное $x_1, x_2, \text{K}\ , x_n$. Совокупность этих значений рассматривается как простая выборка.

Наблюдаемое значение x_i называют *вариантой*, а их последовательность, записанную в возрастающем порядке – *вариационным рядом.* Для каждой варианты можно указать *частоту* ее появления, которую обозначают n_i. Также может быть найдена *относительная частота* появления определенной варианты, как отношение частоты к объему выборки: $w_i = n_i / n$.

Статистическим распределением выборки (*вариационным рядом*) называют соответствие вариант (расположенных в возрастающем порядке) и их частот или относительных частот (табл. 2.1).

Таблица 2.1 Вариационный ряд

Варианта	x_1	x_2	…	x_k	$\sum$
Частота	n_1	n_2	…	n_k	n
Относительная частота	w_1	w_2	…	w_k	1

Если каждую пару (x_i, n_i) изобразить точкой на координатной плоскости и соединить эти точки ломаной линией, то будет получен *полигон частот*. Ломаная, соединяющая точки (x_i, w_i), называется *полигоном относительных частот*. Это аналог многоугольника распределения (рис. 2.1).

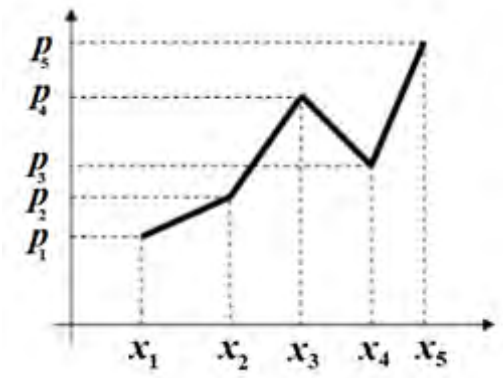


Рис. 2.1 – Полигон относительных частот

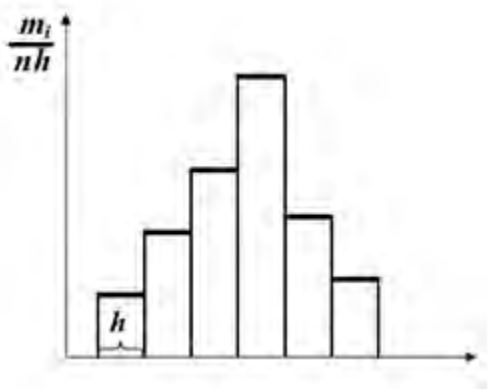


Рис. 2.2 – Гистограмма относительных частот

При большом числе наблюдении статистический ряд перестает быть удобной формой записи статистического материала – он становится громоздким и мало наглядным. Для придания ему большей компактности и наглядности строится так называемый *интервальный статистический (вариационный) ряд* (табл. 2.2). В этом случае весь диапазон наблюдаемых значений X разделяется на интервалы и подсчитывается количество значений n_i, w_i, приходящееся на каждый интервал.

Таблица 2.2 Интервальный статистический ряд

Интервал	$[x_1, x_2)$	$[x_2, x_3)$	…	$[x_{l-1}, x_l]$	$\sum$
Частота	n_1	n_2	…	n_l	n
Относительная частота	w_1	w_2	…	w_l	1

Длину интервала – h – проще выбирать одинаковой. Для нахождения длины интервала можно воспользоваться следующей формулой:

$$h = \frac{R}{l} = \frac{x_{\max} - x_{nin}}{l} \qquad (2.1)$$

Здесь l – количество интервалов, вычисляемое по формуле $l = 1 + 3{,}322 \cdot \log n = \log_2 n + 1$; R – размах выборки.

Если в результате вычисления по формуле (2.1) длина интервала получится дробным числом, то выбирают, либо близкое целое число, либо близкую простую дробь.

По этим данным можно построить гистограммы частот и относительных частот. *Гистограммой* (рис. 2.2) называют ступенчатую фигуру, состоящую из прямоугольников, основаниями которой служат частичные интервалы шириной h, а высоты равны n_i/h (для частот) или w_i/h (для относительных частот).

Гистограмма относительных частот является аналогом функции плотности распределения случайной величины.

Выборочной (эмпирической) функцией распределения называется функция $F^*(x)=n_x/n$, где n_x – число значений случайной величины, меньших x, а n – объем выборки.

При большом числе наблюдений эмпирическая функция распределения приближается к теоретической интегральной функции распределения генеральной совокупности.

Эмпирическая функция распределения обладает всеми свойствами теоретической функции распределения.

Кумулятивная кривая (кумулянта) – ломаная, соединяющая точки с координатами (x_i, n_{x_i}) или $\left(x_i, \frac{n_{x_i}}{n}\right)$, где n_{x_i} – накопленные частоты; для интервального ряда n_{x_i} – число вариант меньших значений вариант i-го интервала.

Накопленная частота (частость) равна сумме всех частот (относительных частот) вариант, предшествующих данному значению. Накопленная частота характеризует число членов данной совокупности, в которых признак, нас интересующий меньше данного значения.

2.2.1 Точечные и интервальные оценки характеристик случайной величины

Важнейшим этапом обработки статистических данных является вычисление оценок числовых характеристик исследуемой случайной величины.

Полученные оценки позволяют в числовой форме описать характерные черты статистического распределения и являются базой для построения математической модели изучаемого случайного явления.

Любая величина $\tilde{\theta}$, определяемая как функция выборочных значений $\tilde{\theta}=\varphi(x_1, x_2, \text{K}, x_n)$, называется *выборочной статистикой* или просто *статистикой*. Статистика $\tilde{\theta}$, используемая в качестве приближённого значения неизвестного параметра θ, называется *статистической оценкой* параметра θ.

Существует два вида оценок параметров: точечные и интервальные. *Точечной* называется статистическая оценка, которая определяется одним числом. К точечным статистическим оценкам предъявляется ряд требований.

Если $\widetilde{\theta}$ – статистическая оценка параметра θ, то она должна удовлетворять следующим условиям:

1) быть *несмещенной*, что означает, что $\mathrm{M}\left(\widetilde{\theta}\right)=\theta$.

2) быть *состоятельной*, т.е. предел по вероятности при $n\to+\infty$ последовательности таких оценок должен быть равен искомому параметру.

3) быть э*ффективной*, т.е. дисперсия $D\left(\widetilde{\theta}\right)$ – наименьшая или быть *асимптотически эффективной*, что означает, что $\lim\limits_{n\to\infty} D\left(\widetilde{\theta}\right)=0$.

Выборочной средней $\bar{x}$ называют среднее арифметическое значение случайной величины X по выборочной совокупности объёма n:

$$\bar{x}=\frac{1}{n}\sum_{i=1}^{k}x_i n_i \qquad (2.2)$$

Выборочная средняя служит *несмещённой оценкой математического ожидания* признака X или генеральной совокупности.

Кроме выборочной средней в статистическом анализе применяются структурные средние: *медиана* и *мода*.

Мода выборки – варианта Mo с наибольшей частотой; медиана Me – варианта, которая делит вариационный ряд на равные части. Если $n=2m+1$, то $Me=x_{m+1}$, а если $n=2m$, то $Me=\dfrac{x_m+x_{m+1}}{2}$.

Средние величины не отражают изменчивости (вариации) значений признака. Чтобы охарактеризовать рассеяние наблюдаемых значений количественного признака выборки вокруг своего среднего значения $\bar{x}$ вводят свободную характеристику – *выборочную дисперсию*.

Выборочной дисперсией D_B называют среднее арифметическое квадратов отклонения наблюдаемых значений признака от их среднего значения $\bar{x}$:

$$D_{\mathrm{B}}=\frac{1}{n}\sum_{i=1}^{k}\left(x_i-\bar{x}\right)^2\cdot n_i=\frac{1}{n}\sum_{i=1}^{k}x_i^2 n_i-\bar{x}^2 \qquad (2.3)$$

Выборочным средним квадратическим отклонением (стандартом) называют квадратный корень из выборочной дисперсии:

$$\sigma_{\mathrm{B}}=\sqrt{D_{\mathrm{B}}} \qquad (2.4)$$

Выборочная дисперсия является смещённой оценкой генеральной дисперсии. В качестве н*есмещенной оценки генеральной дисперсии* служит «исправленная» выборочная дисперсия

$$S^2 = \frac{n}{n-1} D_B = \frac{1}{n-1} \sum_{i=1}^{k} (x_i - \bar{x})^2 \cdot n_i \qquad (2.5)$$

Стандартная ошибка среднего оценивает изменчивость выборочного среднего, приближённо показывая, насколько выборочное среднее отличается от среднего генеральной совокупности: $s_{\bar{x}} = \frac{s}{\sqrt{n}}$.

6) Скошенность кривой называется *асимметрией.* Для выборочной асимметрии $\widetilde{A}_s$ справедлива формула: $\widetilde{A}_s = \frac{1}{n \cdot s^3} \sum_i n_i (x_i - \bar{x})^3$.

7) Отклонение крутизны называют *эксцессом*. Выборочный эксцесс $\widetilde{E}_x$ определяется формулой: $\widetilde{E}_x = \frac{1}{n \cdot s^4} \sum_i n_i (x_i - \bar{x})^4 - 3$.

Так как асимметрия и эксцесс являются характеристиками формы кривой распределения, то по величине выборочных асимметрии и эксцесса можно делать предположения о его виде. Если выборочные асимметрия и эксцесс достаточно малы, т.е. близки к нулю, то можно выдвигать гипотезу о нормальном законе распределения генеральной совокупности.

8) *Коэффициент вариации:* $V = \frac{s}{\bar{x}} \cdot 100\%$.

Коэффициент вариации является относительной мерой рассеяния признака.

Коэффициент вариации используется и как показатель однородности выборочных наблюдений.

Считается, что если коэффициент вариации не превышает 10 %, то выборку можно считать однородной, т. е. полученной из одной генеральной совокупности.

При малом числе наблюдений точечная оценка в значительной степени случайна, и замена истинного значения параметра на оценку может привести к серьезным ошибкам.

Чтобы дать представление о точности и надежности оценки, в математической статистике используют так называемые доверительные интервалы и доверительную вероятность.

Доверительный интервал (confidence interval) – вычисленный на основе выборки интервал значений признака, который с известной вероятностью содержит оцениваемый параметр генеральной совокупности.

Доверительная вероятность (или уровень доверия, confidence level) – это вероятность того, что доверительный интервал содержит значение параметра.

Доверительную вероятность принято устанавливать на уровнях 90, 95 и 99 %. Чем выше доверительная вероятность, тем более широкий и менее полезный получается интервал. Если доверительная вероятность не задана, считают, что она равна 0,95, или 95 %.

Уровень значимости α – это вероятность противоположного события (непопадания истинного значения параметра в доверительный интервал).

Точечной оценкой математического ожидания является выборочное среднее $\bar{x}$. Границы доверительного интервала определяются как $a_1 = \bar{x} - \delta; a_2 = \bar{x} + \delta$, где $\delta > 0$ – точность доверительного интервала, которая либо задается заранее, либо вычисляется.

Предположим, что наблюдается случайная величина *X*, имеющая нормальное распределение с параметрами *a* и σ. Для параметров строятся следующие доверительные интервалы.

1. Для неизвестного среднего a при известной дисперсии σ^2:

$$\bar{x} - t \cdot \frac{\sigma}{\sqrt{n}} < a < \bar{x} + t \cdot \frac{\sigma}{\sqrt{n}}. \qquad (2.6)$$

Значение t находится из соотношения $\Phi(t) = \frac{\gamma}{2}$, где $\Phi(t)$ – функция Лапласа.

2. Для неизвестного среднего *a* при неизвестной дисперсии:

$$\bar{x} - t_\gamma \cdot \frac{s}{\sqrt{n}} < a < \bar{x} + t_\gamma \cdot \frac{s}{\sqrt{n}}. \qquad (2.7)$$

где t_γ – критическая точка распределения Стьюдента (для двусторонней критической области) с числом степеней свободы $k = n - 1$ и уровнем значимости α; s – «исправленное» стандартное отклонение.

Для вычисления критической точки распределения Стьюдента в MS Excel можно воспользоваться следующими функциями:

а) =СТЬЮДЕНТ.ОБР.2Х(α; n–1) – для двусторонней критической области;

б) =СТЬЮДЕНТ.ОБР(1–α; n–1) – для односторонней критической области.

В пакете *Statistica* все необходимые расчеты можно выполнить, используя вероятностный калькулятор.

3. Для неизвестной дисперсии нормально распределенной генеральной совокупности:

$$\frac{(n-1)\cdot s^2}{\chi^2\left(\frac{\alpha}{2},n-1\right)} < \sigma^2 < \frac{(n-1)\cdot s^2}{\chi^2\left(1-\frac{\alpha}{2},n-1\right)} \qquad (2.8)$$

Значения $\chi^2\left(\frac{\alpha}{2},n-1\right)$ и $\chi^2\left(1-\frac{\alpha}{2},n-1\right)$ находятся:

– в пакете *Statistica* – с помощью вероятностного калькулятора;

– в пакете *Excel* – с помощью стандартной функции ХИ2.ОБР.ПХ(α; *k*), где *k – число степеней свободы.*

Иногда полученная точность не удовлетворяет пользователя, так как дает слишком широкий диапазон, в который попадает математическое ожидание с вероятностью *p*. Чем меньше точность доверительного интервала, тем ближе выборочная оценка к соответствующему генеральному показателю.

Точность зависит от числа наблюдений. Можно определить число наблюдений, которые необходимы для достижения заданной точности δ, по следующей формуле:

$$n(\delta) \geq t(\alpha,n-1)\cdot\frac{s^2}{\delta^2}+1. \qquad (2.9)$$

РАЗДЕЛ III Проверка статистических гипотез и дисперсионный анализ

ТЕМА 3.1 Статистическая проверка истинности выдвинутой гипотезы

3.1.1 Общие сведения о проверке статистических гипотез

При обработке экспериментальных данных, при решении многих практических задач для характеристики свойств наблюдаемых случайных величин (СВ) и для проведения теоретических выкладок приходится делать предположения о виде законов распределения этих величин (нормальном, показательном, равномерном и т.д.) или о соотношении между параметрами распределений. Такие предположения называются *гипотезами*. Приняв гипотезу, из нее получают определенные теоретические данные и проверяют, насколько они согласуются с результатами опыта.

Выбор распределения по опытным данным может быть сделан из следующих соображений:

- исходя из физической природы исследуемого объекта;
- по виду гистограммы или полигона частот;
- по опытным данным ранее проведённых исследований;
- с помощью графического представления эмпирической функции;
- с помощью критериев согласия и т.д.

Статистической гипотезой называется любое предположение относительно генеральной совокупности. Гипотеза называется *параметрической*, если в ней содержится некоторое утверждение о параметрах распределения случайной величины (когда сам закон распределения считается известным), и *непараметрической* – в иных случаях.

Нулевой (основной) гипотезой H_0 называется предположение, которого мы придерживаемся изначально, пока наблюдения не заставят нас признать обратное.

Альтернативной (конкурирующей) гипотезой H_1 называется гипотеза, которая противоречит H_0, и которую мы принимаем, если отвергаем основную гипотезу.

Случайная величина K, построенная по наблюдениям для проверки нулевой гипотезы, называется *статистикой критерия.* В каждом конкретном случае статистику критерия подбирают, обычно из следующих: U – нормальное распределение, χ^2 – распределение хи-квадрат (Пирсона), t – распределение Стьюдента, F – распределение Фишера-Снедекора.

Схема построения критерия такова: все выборочное пространство делится на две взаимодополняющие области: область отклонения основной гипотезы H_0 и область принятия этой гипотезы. Область, при попадании в которую выборочной точки отвергается основная гипотеза, называется *критической.*

При проверке гипотезы H_0 возможны следующие ошибки:

• *ошибка первого рода* – отвергнуть гипотезу H_0 при её правильности. Вероятность допустить ошибку первого рода называется *уровнем значимости* α;

• *ошибка второго рода* – принятие гипотезы H_0 при правильности альтернативной гипотезы.

Вероятность принять верную гипотезу называется *уровнем доверия* $\gamma = 1 - \alpha$.

Вероятность принять альтернативную гипотезу, если она верна, называется *мощностью критерия*.

Вычисленное по выборке значение критерия называют *наблюдаемым значением* $K_{\text{набл}}$.

Критическими точками (*границами*) называют точки $k_{\text{кр}}$, отделяющие критическую область от области принятия гипотезы. Критические точки разделяются на правосторонние и левосторонние области. *Правосторонняя* область определяется неравенством $K > k_{\text{кр}}$, *левосторонняя* – $K < k_{\text{кр}}$. Это односторонние области.

Существуют также и двусторонние области, определяемые неравенствами $K < k_{1\text{кр}}$, $K > k_{2\text{кр}}$, где $k_{2\text{кр}} > k_{1\text{кр}}$ ($k_{1\text{кр}}$ и $k_{2\text{кр}}$ – критические точки). Для

каждого критерия, т.е. соответствующего распределения, обычно составлены таблицы, по которым находят $k_{кр}$ (для нахождения критических точек можно использовать стандартные функции математически пакетов).

После того как критическая точка найдена, по данным выборки вычисляют наблюдаемое значение критерии. Если $K_{набл} > k_{кр}$ (для правосторонней области) нулевую гипотезу отвергают, если наоборот, то принимают.

Проверку нулевой гипотезы можно проводить с помощью так называемой *статистической значимости*. Статистическую значимость находят с помощью *p*-значения, которое соответствует вероятности данного события при предположении, что некоторое утверждение (нулевая гипотеза) истинно. Если *p*-значение меньше заданного уровня статистической значимости (обычно это 0,05) – нулевая гипотеза неверна, поэтому нужно перейти к рассмотрению альтернативной гипотезы.

3.1.2 Оценка соответствия выборочных данных теоретическому закону распределения

Пусть $x_1, x_2, ..., x_n$ – выборка наблюдений случайной величины X с неизвестной функцией распределения $F(x)$. Проверяется гипотеза H_0, утверждающая, что X распределена по закону, имеющему функцию распределения $F(x)$, равную функции $F_0(x)$, т.е. проверяется нулевая гипотеза $H_0 : F(x) = F_0(x)$. Критерии, с помощью которых проверяется нулевая гипотеза о неизвестном распределении, называются *критериями согласия*. Рассмотрим *критерий согласия Пирсона* (хи-квадрат распределения).

Схема проверки нулевой гипотезы $H_0 : F(x) = F_0(x)$:

1. По выборке $x_1, x_2, ..., x_n$ строят вариационный ряд; он может быть как дискретным, так и интервальным.

2. По данным предыдущих исследований или по предварительным данным делают предположение (принимают гипотезу) о модели закона распределения случайной величины X.

3. По выборочным данным проводят оценку параметров выбранной модели закона распределения. Предположим, что закон распределения имеет r параметров (например, биномиальный закон имеет один параметр p; нормальный – два параметра (a, σ) и т.д.)

4. Подставляя выборочные оценки значений параметров распределения, находят *теоретические значения вероятностей* $p_i = P(X = x_i)$.

5. Рассчитывают *теоретические частоты* $n = n \cdot p_i$, где n – объем выборки.

6. Рассчитывают значение критерия согласия Пирсона

$$\chi^2 = \sum_{i=1}^{l} \frac{(n_i - n_i')^2}{n_i'} \qquad (3.1)$$

Здесь n_i – частоты данного статистического распределения, n_i' – теоретические частоты, найденные с помощью функции распределения предполагаемого закона;

Эта величина при $n \to \infty$ стремится к распределению χ^2 с $k = l - r - 1$ степенями свободы, где l – число интервалов для интервального вариационного ряда или число групп для дискретного ряда, r – число параметров предполагаемого распределения. В частности, если предполагаемое распределение является нормальным, то оценивается два параметра, поэтому число степеней свободы $k = l - 3$.

7. Задавая уровень значимости α, находят критическую область: она всегда правосторонняя – $(\chi^2_{\text{кр}}; \infty)$; значение $\chi^2_{\text{кр}}$ определяют из соотношения $\alpha = P(\chi^2 > \chi^2_{\text{кр}})$. Если численное значение $\chi^2_{\text{набл}}$ попадает в интервал $(\chi^2_{\text{кр}}; \infty)$, то гипотеза $H_0 : F(x) = F_0(x)$ отклоняется и принимается альтернативная гипотеза о том, что выбранная модель закона распределения не подтверждается выборочными данными, при этом допускается ошибка, вероятность которой равна α.

Критерий согласия Пирсона можно использовать только в том случае, когда $n \cdot p_i \geq 5$. Поэтому тот интервал, для которого это условие не выполняется, объединяют с соседним и соответственно уменьшают число интервалов.

Замечание 3.1. Критическое значение статистики (3.1) можно найти:

– в пакете Excel с помощью стандартной функции ХИ2.ОБР.ПХ(α, k), где α – уровень значимости; k – число степеней свободы;

– в пакете Mathcad с помощью стандартной функции $qchisq(1 - \alpha, k)$.

Замечание 3.2. В качестве меры близости эмпирического и теоретического распределений В.И. Романовский предложил использовать величину χ^2, но с учетом числа степеней свободы k: $c = \dfrac{\left|\chi^2 - k\right|}{\sqrt{2k}}$.

Если величина этого выражения меньше 3, т.е. $c < 3$, то это дает основание для проверки гипотезы H_0, в противном случае, когда $c > 3$, расхождения считаются существенными и гипотеза H_0 о нормальном законе не принимается.

Замечание 3.3. В практике часто используется *приближенная проверка на нормальность*, в основе которой лежит более простые рекомендации, использующие значения числовых характеристик и свойства нормального распределения – известно, что если случайная величина подчиняется нормальному закону распределения, то ее значения удовлетворяют следующим условиям:

– промежуток $\bar{x} \pm 0{,}3\sigma_B$ содержит примерно $\frac{1}{4}$ часть всей совокупности значений;

– промежуток $\bar{x} \pm 0{,}7\sigma_B$ содержит примерно $\frac{1}{2}$ часть;

– промежуток $\bar{x} \pm 1{,}1\sigma_B$ содержит примерно $\frac{3}{4}$ часть;

– промежуток $\bar{x} \pm 3\sigma_B$ содержит примерно 0,99 всех значений.

Если эти соотношения выполняются одновременно для данной эмпирической совокупности и вычисленных $\bar{x}, \sigma_B$, то гипотеза о нормальном законе распределения может быть принята.

Критерий Колмогорова предназначен для проверки гипотезы о законе распределения только непрерывных случайных величин. Он позволяет сравнить эмпирическую функцию $F^*(x)$ и теоретическую функцию распределения $F(x)$.

Схема применения критерия Колмогорова:

1) Для предполагаемого закона распределения нужно определить $F(x)$ для значений аргументов, соответствующих правым концам интервалов.

2) Вычислить значение статистики $\lambda = \sqrt{n} \cdot \max_{x_i} \left| F(x_i) - F^*(x_i) \right|$.

3) По уровню значимости α из таблицы 3.1 найти критическую точку $\lambda_{кр}$. Если $\lambda < \lambda_{кр}$, то различия между эмпирическим и предполагаемым теоретическим распределениями несущественны. Если $\lambda > \lambda_{кр}$, то различия между эмпирическим и предполагаемым теоретическим распределениями существенны.

Таблица 3.1

α	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.001
$\lambda_{кр}$	1.138	1.2238	1.3581	1.4802	1.5174	1.6276	1.738	1.9495

ТЕМА 3.2 Проверка параметрических гипотез

Статистическая гипотеза, которая выдвигает предположение относительно значений параметров функции распределения определённого вида, называется *параметрической*.

3.2.1. Проверка гипотезы о математическом ожидании нормально распределённой случайной величины при неизвестной дисперсии.

Пусть случайная величина $X \sim N(a, \sigma)$, среднее квадратическое отклонение σ и математическое ожидание a – неизвестны. Есть основания предполагать, что $a = a_0$. Тогда $H_0 : a = a_0$; $H_1 : a \neq a_0 \, (a < a_0; a > a_0)$.

Для проверки нулевой гипотезы извлекается выборка объёма n. В качестве критерия выбирается статистика:

$$T = \frac{\bar{x} - a_0}{s} \cdot \sqrt{n} \qquad (3.2)$$

которая при справедливости H_0 имеет распределение Стьюдента с $k = n - 1$ степенями свободы.

Для того чтобы при заданном уровне значимости α проверить $H_0 : a = a_0$ при альтернативной гипотезе $H_1 : a > a_0$, по таблице распределения Стьюдента находят квантили $t_{\text{кр}} = t(\alpha; k)$ из равенства $P(T > t(\alpha, k)) = \alpha$.

Если $T_{\text{набл}} > t_{\text{кр}}$, то нулевая гипотеза отвергается на уровне значимости α; в противном случае нет оснований отвергнуть нулевую гипотезу.

При альтернативной гипотезе $H_1 : a < a_0$, по таблице распределения Стьюдента находят квантиль $t_{\text{кр}} = t(\alpha; k)$ из равенства $P(T > t(\alpha, k)) = 1 - \alpha$.

Если $T_{\text{набл}} \le t_{\text{кр}}$, то нулевая гипотеза отвергается на уровне значимости α; в противном случае нет оснований отвергнуть нулевую гипотезу.

При альтернативной гипотезе $H_1 : a \ne a_0$, сравнивают модуль статистической характеристики T с квантилем $t_{\text{кр}} = t(\alpha; k)$ распределения Стьюдента, найденным из равенства $P\left(T \le t\left(\frac{\alpha}{2}; k\right)\right) = P\left(T \ge t\left(\frac{\alpha}{2}; k\right)\right) = \frac{\alpha}{2}$

Если $|T_{\text{набл}}| < t_{\text{кр}}$, то нет оснований отвергнуть нулевую гипотезу, в противном случае нулевая гипотеза отвергается на уровне значимости α.

Замечание 3.4. В пакете MS Excel квантиль распределения Стьюдента можно найти с помощью стандартных функций СТЬЮДЕНТ.ОБР.2Х(α; k) для двусторонней критической области и СТЬЮДЕНТ.ОБР(1–α; k) – для односторонней; в пакете Statistica – с помощью вероятностного калькулятора; в пакете Mathcad – с помощью функции $qt\left(1 - \frac{\alpha}{2}, k\right)$ (или $qt(1 - \alpha, k)$).

3.2.2. Проверка гипотезы о дисперсии случайной величины X, распределённой по нормальному закону.

Дисперсия характеризует такие важные технологические и конструкторские показатели, как точность машин, погрешность показаний контрольно-измерительных приборов, ритмичность производства, устойчивость работы автоматических линий и др.

Пусть случайная величина X распределена по нормальному закону. Генеральная дисперсия не известна, то есть основания по теоретическим предположениям или по предыдущим опытам считать ее равной σ_0^2. Из генераль-

ной совокупности производится выборка объемом n и вычисляется «исправленная» выборочная дисперсия s^2. Чтобы при заданном уровне значимости α проверить основную гипотезу H_0 о равенстве генеральной дисперсии σ^2 значению σ_0^2 применяется статистика

$$\chi^2 = \frac{n-1}{\sigma_0^2} \cdot s \qquad (3.3)$$

которая при справедливости гипотезы H_0 имеет распределение Пирсона с $n-1$ степенями свободы.

Возможны три случая выдвижения альтернативной гипотезы:

1. $H_1 : \sigma^2 > \sigma_0^2$. В этом случае критическая область ищется, как правосторонняя из условия $P\left(\chi^2 > \chi^2_{\text{кр}}(\alpha; k)\right) = \alpha$, а критическую точку ищут по таблицам квантилей распределения χ^2 (или с помощь стандартных функций математических пакетов). После этого вычисляем по данной выборке наблюдаемое значение критерия. Если $\chi^2_{\text{набл}} < \chi^2_{\text{кр}}(\alpha; k)$, то нулевая гипотеза принимается.

2. $H_1 : \sigma^2 < \sigma_0^2$. В этом случае критическую область ищут как левостороннюю. Критическая точка ищется как $\chi^2_{\text{кр}}(1-\alpha; k)$. Тогда, если $\chi^2_{\text{набл}} > \chi^2_{\text{кр}}(\alpha; k)$, то нулевая гипотеза принимается.

3. $H_1 : \sigma^2 \neq \sigma_0^2$. В этом случае критическая область ищется как двусторонняя. Критические точки находятся из условий:

$$P\left(\chi^2 < \chi^2_{\text{лев}}\left(1-\frac{\alpha}{2}\right)\right) = \frac{\alpha}{2}; P\left(\chi^2 > \chi^2_{\text{прав}}\left(\frac{\alpha}{2}; k\right)\right) = \frac{\alpha}{2}.$$

Если $\chi^2_{\text{лев}} < \chi^2_{\text{набл}} < \chi^2_{\text{прав}}$ – нет оснований отвергнуть нулевую гипотезу. Если $\chi^2_{\text{набл}} < \chi^2_{\text{лев}}$ или $\chi^2_{\text{набл}} > \chi^2_{\text{прав}}$ – нулевую гипотезу отвергают.

3.2.3. Проверка гипотезы о дисперсиях двух случайных величин, распределённых по нормальному закону.

Задача сравнения дисперсий возникает при сравнении точности приборов, инструментов и др. Прибор, который обеспечивает наименьшую дисперсию, является лучшим.

Пусть исследуются 2 случайные величины X и Y, распределённые по нормальному закону с неизвестными параметрами (a_1, σ_1) и (a_2, σ_2). Из генеральных совокупностей выполнены выборки объёмами n_1 и n_2, и вычислены точечные оценки $\bar{x}, \bar{y}, s_x^2, s_y^2$. Выдвигается нулевая гипотеза, состоящая

в том, что генеральные дисперсии рассматриваемых совокупностей равны между собой: $H_0 : \sigma_1^2 = \sigma_2^2$.

Для проверки нулевой гипотезы вычисляется наблюдаемое значение критерия

$$F_{\text{набл}} = \frac{s_{\text{Б}}^2}{s_{\text{М}}^2} \qquad (3.4)$$

где $s_{\text{Б}}^2$, $s_{\text{М}}^2$ – соответственно большая и меньшая «исправленные» дисперсии.

Случайная величина F имеет распределение Фишера с $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$ степенями свободы. Критическая область строится в зависимости от вида конкурирующей гипотезы.

1) $H_0 : \sigma_1^2 = \sigma_2^2;\ H_1 : \sigma_1^2 \neq \sigma_2^2$.

В этом случае строят двустороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия в эту область, в предположении справедливости нулевой гипотезы, была равна принятому уровню значимости α. Наблюдаемое значение критерия вычисляется по формуле (3.4).

Если $F_{\text{набл}} < F_{\text{кр}}$, то гипотеза о равенстве дисперсий принимается. Если $F_{\text{набл}} > F_{\text{кр}}$ – нулевую гипотезу отвергают.

2) $H_0 : \sigma_1^2 = \sigma_2^2;\ H_1 : \sigma_1^2 > \sigma_2^2$.

В этом случае строят правостороннюю критическую область, исходя из требования, чтобы вероятность попадания критерия F в эту область в предположении справедливости нулевой гипотезы была равна принятому уровню значимости: $P\left[F > F_{\text{кр}}(\alpha; k_1; k_2)\right] = \alpha$.

Наблюдаемое значение критерия вычисляется по формуле (3.4).

Если $F_{\text{набл}} < F_{\text{кр}}$, то гипотеза о равенстве дисперсий принимается. Если $F_{\text{набл}} > F_{\text{кр}}$ – нулевую гипотезу отвергают.

Критическое значение статистики (3.4) можно найти:

– в пакете Excel с помощью стандартной функции F.ОБР.ПХ(α, k_1, k_2), где α – уровень значимости; k_1 – число степеней свободы большей дисперсии;

– в пакете Mathcad с помощью стандартной функции $qF(1-\alpha, k_1, k_2)$ для односторонней критической области; $qF\left(1-\frac{\alpha}{2}, k_1, k_2\right)$ – для двусторонней.

3.2.4. Проверка гипотез о равенстве математических ожиданий двух случайных величин, распределённых по нормальному закону.

Обозначим через n_1 и n_2 объёмы малых независимых выборок, по которым найдены соответствующие выборочные средние $\bar{x}$ и $\bar{y}$, а также исправленные выборочные дисперсии s_x^2 и s_y^2.

1) Проверяемая гипотеза $H_0 : a_1 = a_2$, дисперсии равны, но неизвестны. Принимается, что оценками σ_x^2 и σ_y^2 являются s_x^2 и s_y^2. Статистикой критерия является величина:

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{(n_1 - 1)s_x^2 + (n_2 - 1)s_y^2}} \cdot \sqrt{\frac{n_1 n_2 (n_1 + n_2 - 2)}{n_1 + n_2}} \qquad (3.5)$$

В том случае, когда проверяемая гипотеза верна, статистика, определяемая формулой (3.5), имеет распределение Стьюдента с $n_1 + n_2 - 2$ степенями свободы. Область принятия гипотезы H_0 для двусторонней критической области (альтернативная гипотеза $H_1 : a_1 \neq a_2$) имеет вид: $|T_{\text{набл}}| < t_{\text{двуст. кр.}}(\alpha; k)$

Здесь

- $T_{\text{набл}}-$ наблюдаемое значение критерия – находится по формуле (3.5);
- $k = n_1 + n_2 - 2$ – число степеней свободы;
- $t_{\text{двуст. кр.}}(\alpha; k)$ – критическая точка двусторонней критической области.

При конкурирующей гипотезе $H_1 : a_1 > a_2$ находят критическую точку $t_{\text{правост.кр.}}(\alpha; k)$ для односторонней критической области.

Если $T_{\text{набл}} < t_{\text{правост.кр.}}(\alpha; k)$ – нет оснований отвергнуть нулевую гипотезу.

Если $T_{\text{набл}} > t_{\text{правост.кр.}}(\alpha; k)$ – нулевую гипотезу отвергают.

Если $H_1 : a_1 < a_2$, то находят сначала критическую точку $t_{\text{правост.кр.}}(\alpha; k)$ и полагают $t_{\text{левост.крит}}(\alpha; k) = -t_{\text{правост.кр.}}(\alpha; k)$. Если $T_{\text{набл}} > -t_{\text{правост.кр.}}(\alpha; k)$ – нет оснований отвергнуть нулевую гипотезу. Если $T_{\text{набл}} < t_{\text{правост.кр.}}(\alpha; k)$ – нулевую гипотезу отвергают.

2) Если дисперсии генеральных совокупностей неизвестны и не предполагаются равными, то можно приближённо считать, что статистика

$$T = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}} \qquad (3.6)$$

также подчинена распределению Стьюдента. Но число степеней свободы уже не является целым числом:

$$k = \frac{\left(\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}\right)^2}{\frac{\left(\frac{s_x^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_y^2}{n_2}\right)^2}{n_2 - 1}} \qquad (3.7)$$

Область принятия гипотезы для двусторонней критической области (альтернатива $H_1 : a_1 \neq a_2$) имеет вид:

$$|T| < t_{\text{двуст.крит.}}\left(\frac{\alpha}{2}; k\right) \qquad (3.8)$$

3.2.5. Сравнение двух средних нормальных генеральных совокупностей с неизвестными дисперсиями (зависимые выборки)

Пусть генеральные совокупности X и Y распределены нормально, причём их дисперсии неизвестны. Из этих совокупностей извлечены зависимые выборки одинакового объёма n, варианты которых соответственно равны x_i и y_i. Введём следующие обозначения:

$d_i = x_i - y_i$ – разности вариант с одинаковыми номерами,

$\overline{d} = \frac{\sum\limits_i d_i}{n}$ – средняя разностей вариант с одинаковыми номерами;

$s_d = \sqrt{\frac{\sum\limits_i d_i^2 - \frac{\left(\sum\limits_i d_i\right)^2}{n}}{n-1}}$ «исправленное» среднее квадратическое отклонение.

Для того чтобы при заданном уровне значимости α проверить нулевую гипотезу $H_0 : a_1 = a_2$ о равенстве двух средних нормальных совокупностей X и Y с неизвестными дисперсиями (в случае зависимых выборок одинакового объёма) при конкурирующей гипотезе $H_1 : a_1 \neq a_2$ нужно:

1) вычислить наблюдаемое значение критерия $T_{\text{набл}} = \frac{\overline{d}}{s_d} \cdot \sqrt{n}$;

2) по таблице критических точек распределения Стьюдента (см. [4]), по заданному уровню значимости α для двусторонней критической области и числу степеней свободы $k = n - 1$ найти критическую точку $t_{\text{двуст.крит.}}(\alpha; k)$;
3) если $|T_{\text{набл}}| < t_{\text{двуст.крит.}}$ – нет оснований отвергать нулевую гипотезу. Если $|T_{\text{набл}}| > t_{\text{двуст.крит.}}$ – нулевую гипотезу отвергают.

3.2.6. Сравнение нескольких дисперсий нормальных генеральных совокупностей по выборкам одинакового объёма. Критерий Кохрена.

Пусть генеральные совокупности $X_1, X_2, \ldots, X_l$ распределены нормально. Из этих совокупностей извлечено l выборок одинакового объёма n и по ним найдены исправленные выборочные дисперсии $s_1^2, s_2^2, \ldots, s_l^2$, все с одинаковым числом степеней свободы $k = n - 1$. Требуется по исправленным дисперсиям при заданном уровне значимости α проверить нулевую гипотезу, состоящую в том, что генеральные дисперсии рассматриваемых совокупностей равны между собой: $D(X_1) = D(X_2) = \ldots = D(X_l)$.

Другими словами, требуется проверить, значимо или незначимо различаются исправленные выборочные дисперсии.

В качестве критерия проверки нулевой гипотезы примем критерий Кохрена – отношение максимальной исправленной дисперсии к сумме всех исправленных дисперсий:

$$G = \frac{s_{\max}^2}{s_1^2 + s_2^2 + \ldots + s_l^2}. \qquad (3.9)$$

Распределение этой случайной величины зависит только от числа степеней свободы $k = n - 1$ и количества выборок l.

Критическую точку строят правостороннюю, исходя из требования, чтобы вероятность попадания критерия в эту область, в предположении справедливости нулевой гипотезы, была равна принятому уровню значимости $P\left(G > G_{\text{кр}}(\alpha, k, l)\right) = \alpha$.

Критическую точку $G_{\text{кр}}(\alpha; k; l)$) находят по таблице (см. например [4]), или с помощью стандартной функции пакета Excel БЕТА.ОБР $\left(1 - \frac{\alpha}{l}; \frac{n-1}{2}; \frac{l \cdot (n-1)}{2}\right)$. Тогда правосторонняя критическая область определяется неравенством $G > G_{\text{кр}}$, а область принятия нулевой гипотезы – $G < G_{\text{кр}}$.

При условии однородности дисперсий независимых выборок одинакового объема в качестве оценки генеральной дисперсии принимают среднюю арифметическую исправленных дисперсий.

ТЕМА 3.3 Дисперсионный анализ

Задачей дисперсионного анализа является изучение одного или нескольких факториальных признаков (факторов) на результативный признак (наблюдаемую случайную величину).

Например, если измерения некоторой величины проводятся на k различных приборах, то можно исследовать влияние фактора «прибор» на результаты измерений, т.е. ответить на вопрос, имеют ли различные приборы одну и ту же систематическую ошибку (проверяется гипотеза о равенстве средних). По числу факторов, влияние которых исследуется, различают однофакторный и многофакторный дисперсионный анализ.

Однофакторный дисперсионный анализ используется в тех случаях, когда есть в распоряжении три или более независимые выборки, полученные из одной генеральной совокупности путем изменения какого-либо независимого фактора, для которого по каким-либо причинам нет количественных измерений.

Для этих выборок предполагают, что они имеют разные выборочные средние и одинаковые выборочные дисперсии.

Предположим, что на количественный признак X воздействует фактор F, который имеет несколько градаций (уровней, групп). Для каждого уровня зафиксирована выборка значений. Причём в общем случае размеры этих выборок могут быть различны.

Таким образом, имеем несколько случайных величин $X_1, X_2, ..., X_m$, где m – число уровней фактора. Каждая случайная величина X_j соответствует определённому уровню фактора F_j и для неё получена выборка значений $\{x_{1j}, x_{2j}, ...x_{nj}\}$, где n_j – число наблюдений для данного уровня. Данные наблюдений можно представить в виде таблицы 3.2, в которой количество элементов в столбце может быть различным. При этом $n = n_1 + n_2 + ... + n_m$ – общее число всех наблюдений.

Требуется при уровне значимости α проверить гипотезу о равенстве математических ожиданий, соответствующих уровням:

$$H_0 : M(X_1) = M(X_2) = ... = M(X_m).$$

Другими словами, требуется установить, значимо или незначимо различаются групповые средние.

Суть дисперсионного анализа состоит в сравнении дисперсии, которая обусловлена случайными причинами, с дисперсией, вызванной влиянием исследуемого фактора. Если они значимо различаются, то считают, что фактор оказывает влияние на исследуемую величину. Тогда и математические ожидания для уровней будут различаться. Иногда дисперсионный анализ применяют, чтобы установить однородность нескольких совокупностей. Однород-

ные совокупности можно объединить в одну и тем самым получить о ней более полную информацию и более надёжные выводы.

Таблица 3.2

Номера наблюдений	Уровни (группы) фактора			
	F_1	F_2	…	F_m
1	x_{11}	x_{12}	…	x_{1m}
2	x_{21}	x_{22}	…	x_{2m}
3	x_{31}	x_{32}	…	x_{3m}
…	…	…	…	…
n_j	$x_{n_1 1}$	$x_{n_2 2}$	…	$x_{n_m m}$
Групповая средняя $\bar{x}_j$	$\bar{x}_1$	$\bar{x}_2$		$\bar{x}_m$

Дисперсионный анализ может быть применён, если:

1) генеральные совокупности $X_1, X_2, ..., X_m$ распределены нормально и имеют одинаковую, хотя и неизвестную, дисперсию;

2) наблюдения независимы и проводятся в одинаковых условиях.

Проверка нулевой гипотезы основана на сопоставлении двух оценок неизвестной дисперсии σ^2. Обозначим:

$$\bar{x}_j = \frac{1}{n_j}\sum_{i=1}^{n_j} x_{ij} \quad \left(j = \overline{1,m}\right) - \textit{групповые средние}$$

$$\bar{x} = \sum_{j=1}^{m}\sum_{i=1}^{n_j} n_{ij} \quad - \textit{общая выборочная средняя}.$$

Несмещённой оценкой для неизвестной дисперсии σ^2 является сумма квадратов $\sum_{j=1}^{m}\sum_{i=1}^{n_j}\left(x_{ij} - \bar{x}\right)^2$, делённая на $n-1$, где $n = \sum_{j=1}^{m} n_j$ – количество всех наблюдений (если на каждом уровне проведено одинаковое количество наблюдений $n_1 = n_2 = ... = n_m = n'$, то $n = n' \cdot m$). Основная идея дисперсионного анализа заключается в разбиении этой суммы квадратов отклонений на несколько компонент, каждая из которых соответствует предполагаемой причине изменения средних значений $\bar{x}_j$.

Обозначим:

$Q_{\text{общ}} = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}\right)^2$ – *общая сумма квадратов отклонений* наблюдаемых значений от общей средней;

$Q_{\text{факт}} = \sum_{j=1}^{m} \left(\bar{x}_j - \bar{x}\right)^2 n_i$ – *факторная (межгрупповая) сумма квадратов отклонений* групповых средних от общей средней, которая характеризует рассеяние «между группами» и отражает влияние фактора;

$Q_{\text{ост}} = \sum_{j=1}^{m} \sum_{i=1}^{n_j} \left(x_{ij} - \bar{x}_j\right)^2$ – *остаточная (внутригрупповая) сумма квадратов отклонений* наблюдаемых значений группы от своих групповых средних, которая характеризует рассеяние «внутри группы» и отражает влияние случайных причин.

Справедливо *основное тождество* дисперсионного анализа:

$$Q_{\text{общ}} = Q_{\text{факт}} + Q_{\text{ост}}. \tag{3.10}$$

В дисперсионном анализе анализируются не сами суммы квадратов отклонений, а так называемые средние квадраты, являющиеся несмещёнными оценками соответствующих дисперсий, которые получаются делением сумм квадратов отклонений на соответствующее число степеней свободы. Число степеней свободы определяется как общее число наблюдений минус число связывающих их уравнений. Поэтому несмещённой оценкой *межгрупповой (факторной)* дисперсии является $s^2_{\text{факт}} = \dfrac{Q_{\text{факт}}}{m-1}$, так как при расчёте $Q_{\text{факт}}$ используется m групповых средних, связанных между собой одним уравнением. Несмещённой оценкой внутригрупповой (остаточной) дисперсии является $s^2_{\text{ост}} = \dfrac{Q_{\text{ост}}}{n-m}$, ибо при расчёте $Q_{\text{ост}}$ используются все n наблюдений, связанных между собой m уравнениями $\bar{x}_j = \dfrac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}$. В случае однофакторного комплекса $s^2_{\text{факт}}$ и $s^2_{\text{ост}}$ являются несмещёнными и независимыми оценками дисперсии σ^2.

Сравним обе оценки $s^2_{\text{факт}}$ и $s^2_{\text{ост}}$. Если гипотеза H_0 верна, то *дисперсионное отношение* (статистика):

$$F = \frac{s^2_{\text{факт}}}{s^2_{\text{ост}}} \tag{3.11}$$

имеет распределение Фишера с $k_1 = m - 1$ и $k_2 = n - m$ степенями свободы.

Гипотеза H_0 отвергается, если фактически вычисленное значение статистики F больше критического $F_{\text{кр}}(\alpha, k_1, k_2)$ и принимается, если $F < F_{\text{кр}}(\alpha, k_1, k_2)$.

Степень влияния фактора на результативный показатель может быть измерена с помощью выборочного коэффициента детерминации: $R^2 = \dfrac{Q_{\text{факт}}}{Q_{\text{ост}}}$, показывающего, какова доля общей вариации объясняется влиянием исследуемого фактора.

Замечание 3.5. Если факторная и остаточная дисперсии различаются незначимо, то влияние фактора можно считать незначительным и, следовательно, принять гипотезу о равенстве математических ожиданий.

Замечание 3.6. Границу правосторонней критической области можно найти, используя:

- в пакете *Excel* – функцию F.ОБР.ПХ (α, k_1, k_2);
- в пакете Mathcad – функцию $qF(1-\alpha, k_1, k_2)$.

Двухфакторным дисперсионным анализом называют метод, проверяющий влияние двух независимых переменных (факторов) на зависимую переменную. Кроме этого, исследуется эффект взаимодействия между двумя независимыми переменными.

Для применения метода необходимо выполнение нескольких *условий:*

1) Генеральные совокупности, из которых извлечены выборки, имеют нормальное распределение.

2) Выборки независимы.

3) Дисперсии генеральных совокупностей равны.

4) Выборки (группы) имеют одинаковый объем.

А) Двухфакторный дисперсионный анализ без повторений.

Рассмотрим случайную величину *X*, на которую воздействуют два фактора: *A* и *B*.

Предполагается, что взаимодействие между факторами *A* и *B* отсутствует, а их воздействие может повлиять только на среднее m случайной величины *X*, но никак не влияет на ее дисперсию σ^2. Пусть *a* – число групп фактора *A* и *b* – число групп фактора *B*. Сумма квадратов остатков разделяется на три компоненты: $Q = Q_A + Q_B + Q_e$, где

$Q = \sum_{i=1}^{a} \sum_{j=1}^{b} (x_{ij} - \bar{x})^2$ – общая сумма квадратов отклонений;

$Q_A = b \cdot \sum_{i=1}^{a} (\bar{x}_i - \bar{x})^2$ – объяснённая влиянием фактора A сумма квадратов отклонений;

$Q_B = a \cdot \sum_{j=1}^{b} (\bar{x}_j - \bar{x})^2$ – объяснённая влиянием фактора B сумма квадратов отклонений;

$Q_e = \sum_{i=1}^{a} \sum_{j=1}^{b} (x_{ij} - \bar{x}_i - \bar{x}_j + \bar{x})^2$ – необъяснённая сумма квадратов отклонений или сумма квадратов отклонений ошибки;

$\bar{x} = \frac{1}{n} \sum_{i=1}^{a} \sum_{j=1}^{b} x_{ij}$ – общее среднее наблюдений;

$\bar{x}_i = \frac{1}{b} \sum_{j=1}^{b} x_{ij}$ – среднее число наблюдений в каждой группе фактора A;

$\bar{x}_j = \frac{1}{a} \sum_{i=1}^{a} x_{ij}$ – среднее число наблюдений в каждой группе фактора B.

Дисперсии вычисляются следующим образом:

$s_a^2 = \frac{Q_A}{a-1}$ – дисперсия, объяснённая влиянием фактора A;

$s_b^2 = \frac{Q_B}{b-1}$ – дисперсия, объяснённая влиянием фактора B;

$s_e^2 = \frac{Q_e}{(a-1) \cdot (b-1)}$ – необъяснённая дисперсия или дисперсия ошибки,

причем

$k_A = a - 1$ – число степеней свободы дисперсии, объяснённой влиянием фактора A;

$k_B = b - 1$ – число степеней свободы дисперсии, объяснённой влиянием фактора B;

$k_e = (a-1)(b-1)$ – число степеней свободы необъяснённой дисперсии или дисперсии ошибки;

$k = ab - 1$ – общее число степеней свободы.

Если факторы не зависят друг от друга, то для определения существенности факторов выдвигаются две нулевые гипотезы и соответствующие альтернативные гипотезы:

для фактора A:

$H_0 : m_{1A} = m_{2A} = \text{K} = m_{\alpha A}$; H_1 : не все m_{iA} равны;

для фактора B:

$H_0 : m_{1B} = m_{2B} = \text{К} = m_{\alpha B}$; H_1 : не все m_{jB} равны.

Чтобы определить влияние фактора A нужно наблюдаемое отношение Фишера $F_a = \dfrac{s_a^2}{s_e^2}$ сравнить с критическим отношением Фишера $F_{\text{кр}}(\alpha; k_A; k_e)$.

Чтобы определить влияние фактора B, нужно фактическое отношение Фишера $F_b = \dfrac{s_b^2}{s_e^2}$ сравнить с критическим отношением Фишера $F_{\text{кр}}(\alpha; k_B; k_e)$.

Если фактическое отношение Фишера больше критического отношения Фишера, то следует отклонить нулевую гипотезу с уровнем значимости α. Это означает, что фактор существенно влияет на данные: данные зависят от фактора с вероятностью $\gamma = 1 - \alpha$.

Если фактическое отношение Фишера меньше критического отношения Фишера, то следует принять нулевую гипотезу с уровнем значимости α. Это означает, что фактор не оказывает существенного влияния на данные с вероятностью $\gamma = 1 - \alpha$.

Б) *Двухфакторный дисперсионный анализ с повторениями*.

Двухфакторный дисперсионный анализ с повторениями применяется для того, чтобы проверить не только возможную зависимость результативного признака от двух факторов – A и B, но и возможное взаимодействие факторов A и B. Тогда a – число групп фактора A и b – число групп фактора B, r – число повторений; n – число наблюдений в каждой группе

В таблице 3.3 приведена схема двухфакторного дисперсионного анализа с повторениями:

Таблица 3.3

	Сумма квадратов	Число степеней свободы	Дисперсия	Критерий Фишера
Фактор A	Q_A	$k_A = a - 1$	$s_A^2 = \dfrac{Q_A}{k_A}$	$F_A = \dfrac{s_A^2}{s_e^2}$
Фактор B	Q_B	$k_B = b - 1$	$s_B^2 = \dfrac{Q_B}{k_B}$	$F_B = \dfrac{s_B^2}{s_e^2}$
Взаимодействие $A \times B$	Q_{AB}	$k_{AB} = (a-1)(b-1)$	$s_{AB}^2 = \dfrac{Q_{AB}}{k_{AB}}$	$F_{AB} = \dfrac{s_{AB}^2}{s_e^2}$
Ошибка	Q_e	$k_e = ab(n-1)$	$s_e^2 = \dfrac{Q_e}{k_e}$	

Здесь:

$$Q_A = \frac{a}{N}\sum_{i=1}^{a} c_i^2 - \frac{c^2}{N}\left(c = \sum x_{ijk}; c_i^2 = \sum_{j=1}^{b}\sum_{k=1}^{r} x_{ijk}; N = a \cdot b \cdot n\right)$$ – сумма квадратов для фактора A;

$$Q_B = \frac{b}{N}\sum_{j=1}^{b} c_j^2 - \frac{c^2}{N}\left(c_j = \sum_{i=1}^{a}\sum_{k=1}^{r} x_{ijk}\right)$$ – сумма квадратов для фактора B;

$$Q_{AB} = \frac{ab}{N}\sum_{i=1}^{a}\sum_{j=1}^{b} c_{ij}^2 - Q_A - Q_B - \frac{c}{N}\left(c_{ij} = \sum_{k=1}^{r} x_{ijk}\right)$$ – сумма квадратов для взаимодействия $A \times B$;

$$Q_A = c_0 - \frac{1}{n}\sum_{i=1}^{a}\sum_{j=1}^{b} c_{ij}^2 \left(c_0 = \sum x_{ijk}^2\right)$$ – остаточная сумма квадратов.

Далее необходимо сравнить полученные *F*-значения с критической областью. Нужно начать проверку с гипотезы о взаимодействии факторов. Затем проверяются последовательно гипотезы о влиянии факторов.

РАЗДЕЛ IV Парный корреляционно-регрессионный анализ и нелинейная регрессия

ТЕМА 4.1 Корреляционный анализ

4.1.1 Основные понятия и определения

Одной из основных задач математической статистики является исследование зависимости между двумя или несколькими переменными (случайными величинами). *Функциональной* называют зависимость, каждому значению случайной величины *X* соответствует единственное значение случайной величины *Y*, задается формулой $y = f(x)$.

Строгая функциональная зависимость реализуется редко, так как одна или обе величины подвержены еще и случайным факторам. *Статистической* (или *стохастической, вероятностной*) называется зависимость, при которой изменение одной из величин влечет за собой изменение распределения другой. В частности, статистическая зависимость проявляется в том, что при изменении одной из величин изменяется среднее значение другой. В этом случае статистическую зависимость называют *корреляционной*.

Корреляционный анализ позволяет на основе выборочных данных оценить наличие, направленность и силу статистической взаимосвязи.

Существует несколько основных практических приемов проведения данного анализа: составление корреляционной таблицы и построение корреляционного поля; вычисление выборочной ковариации (корреляционного момента); выборочных коэффициентов корреляции; проверка значимости связи. Каждый из этих приемов может использоваться в зависимости от вида

корреляционного анализа, которых существует несколько: *выборочная* и *ранговая* корреляции.

Для проверки правильности нахождения корреляционной зависимости при выборочном анализе обычно строят *поле корреляции* (*диаграмму рассеяния*). Оно представляет собой отображение геометрических мест значений исследуемых параметров в прямоугольной системе координат. Корреляционное поле позволяет дать наглядную графическую интерпретацию коэффициента корреляции, по виду поля можно судить о виде корреляционной зависимости между параметрами, т.е. оценить *направленность* статистической зависимости.

4.1.2. Ковариация и корреляция

Пусть для двух показателей X и Y (случайных величин) имеется выборка связанных пар наблюдений $\{(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)\}$, где n – число наблюдений.

Выборочной ковариацией (*корреляционным моментом*) K_{XY} называется величина:

$$K_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y}), \tag{4.1}$$

где $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$; $\bar{y} = \frac{1}{n} \sum_{i=1}^{n} y_i$ – средние значения выборочных данных для величин X и Y соответственно.

Ковариация является мерой зависимости случайных величин. Если ковариация равна нулю, то взаимосвязь величин отсутствует. Если $K_{XY} > 0$, то существует прямая зависимость, а если $K_{XY} < 0$ – обратная. Но эта характеристика обладает рядом существенных недостатков. Во-первых, она не позволяет оценить силу зависимости между ними. Во-вторых, её значение зависит от единиц измерения исследуемых случайных величин. Для устранения данных недостатков вводится относительная мера зависимости (безразмерная величина) – *коэффициент корреляции*.

Рассмотрим коэффициент линейной корреляции (Пирсона), который характеризует степень линейной зависимости двух случайных величин.

Выборочным коэффициентом линейной корреляции r_B случайных величин X и Y называется величина, определяемая по формуле:

$$r_B = \frac{K_{XY}}{s_x s_y} = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{n} (x_i - \bar{x})^2 (y_i - \bar{y})^2}}. \tag{4.2}$$

Здесь s_x, s_y – исправленные СКО.

Коэффициент линейной корреляции всегда удовлетворяет соотношению $-1 \le r_B \le 1$.

Если $r_B = 0$, то линейная взаимосвязь между случайными величинами отсутствует. Это может означать, что данные случайные величины независимы либо между ними существует нелинейная зависимость (например, показательная, логарифмическая или другая).

Если $0 < r_B < 1$, то между X и Y существует прямая линейная зависимость. Это означает, что увеличение одного признака ведёт к увеличению другого. Например, при увеличении температуры возрастает давление газа.

Если $-1 < r_B < 0$ 0, то между X и Y имеется обратная линейная зависимость. Это означает, что увеличение одного признака ведёт к уменьшению другого. Например, связь между температурой воздуха и количеством топлива, расходуемого на обогрев помещения.

Если $r_B = \pm 1$, то между X и Y существует линейная функциональная зависимость.

Степень линейной зависимости можно качественно оценить с помощью шкалы Чаддока (табл. 4.1):

Таблица 4.1

$\lvert r_B \rvert$	0.1 – 0.3	0.3 – 0.5	0.5 – 0.7	0.7 – 0.9	0.9 – 0.99
Теснота связи	Слабая	Умеренная	Заметная	Высокая	Весьма высокая

При исследовании связи между несколькими случайными величинами находят выборочные коэффициенты корреляции между парами всех исследуемых величин и строят корреляционную матрицу.

Корреляционная матрица – это квадратная таблица, в которой на пересечении строки i и столбца j находится коэффициент корреляции r_{ij} между случайными величинами X_i и X_j. Эта матрица является симметричной, поэтому часто указывается только половина таблицы (например, под главной диагональю). По диагонали стоят единицы, так как каждая величина полностью коррелирует сама с собой.

Выборочный коэффициент корреляции обычно используется в предположении нормальности данных. В этом случае из равенства нулю теоретического коэффициента r_{XY} следует независимость случайных величин (в более общем случае это неверно). В случае нормального распределения можно проверить гипотезу $H_0 : r_{XY} = 0$. Пусть

$$T = \frac{r_B \cdot \sqrt{n-2}}{\sqrt{1-r_B^2}}. \qquad (4.3)$$

Если гипотеза H_0 верна, то T имеет распределение Стьюдента с $n-2$ степенями свободы. При уровне значимости α выберем критическую точку $t_{\text{кр}}(\alpha, n-2)$ для двусторонней области. Если $|T| < t_{\text{кр}}$, то гипотеза H_0 принимается, выборочный коэффициент корреляции незначим, величины X и Y не коррелированы; иначе – отвергается.

Оценку коэффициента корреляции в генеральной совокупности можно выполнить путём построения доверительного интервала:

$$r_B - t(\alpha, n-2)\cdot \sigma_{r_B} < r_{XY} < r_B + t(\alpha, n-2)\cdot \sigma_{r_B}. \qquad (4.4)$$

Здесь $\sigma_{r_B} = \frac{1-r_B^2}{\sqrt{n}}$ – среднее квадратичное отклонение выборочного коэффициента корреляции; $t(\alpha, n-2)$ – табличное значение критерия Стьюдента для двусторонней критической области.

Если нуль окажется внутри интервала, то коэффициент корреляции в генеральной совокупности равен нулю и выборочный коэффициент парной корреляции будет несущественным.

4.1.3. Ранговая корреляция.

В случае, когда нормальность данных нарушается, применение выборочного коэффициента корреляции может вести к ошибкам: либо мы «не заметим» зависимость между величинами, либо получим ложную корреляцию. Существуют коэффициенты и методы, свободные от предположения о нормальности.

Наблюдения всегда можно упорядочить по возрастанию какой-либо переменной (x или y). *Рангом наблюдения* называется его номер в таком ряду. Если какое-то значение переменной встречается несколько раз, ему приписывается средний ранг. Обозначим ранги наблюдений по возрастанию x и y через r_i, и s_i соответственно. Пусть $S = \sum_{i=1}^{n} (r_i - s_i)^2$.

Коэффициентом ранговой корреляции Спирмена называется величина

$$\rho_B = 1 - \frac{6 \cdot S}{n^3 - n}. \qquad (4.5)$$

Этот коэффициент также может принимать значения от – 1 до +1. Аналогичным образом он отражает силу и характер зависимости между величи-

нами. Для проверки гипотезы о независимости случайных величин существуют специальные таблицы критических точек. Однако при больших n можно проверять гипотезу так же, как для обычного выборочного коэффициента корреляции.

Если гипотеза о независимости справедлива и $n \to \infty$, то распределение статистики $T = \dfrac{\rho_B \cdot \sqrt{n-1}}{\sqrt{1-\rho_B^2}}$ сходится к распределению Стьюдента с $n-2$ степенями свободы. При $n \geq 10$ эту статистику используют для проверки гипотезы о независимости порядковых переменных. Если рассчитанное значение *t*-критерия меньше табличного (для двусторонней критической области) при заданном числе степеней свободы, статистическая значимость наблюдаемой взаимосвязи – отсутствует. Если больше, то корреляционная связь считается статистически значимой (альтернативная гипотеза предполагает, что рассматриваемые признаки зависимы).

С помощью коэффициента Спирмена можно анализировать также ситуации, когда некоторый признак объекта («качество», «привлекательность» и т.п.) нельзя строго выразить численно, но можно упорядочить объекты по его возрастанию или убыванию, т.е. проранжировать их.

Можно оценить связь между двумя качественными признаками, используя *коэффициент ранговой корреляции Кендалла*. Пусть ранги объектов выборки расположены в таблице 4.2:

Таблица 4.2

для признака *X*	r_1	r_2	…	r_n
для признака *Y*	s_1	s_2	…	s_n

Пусть справа от s_1 имеется $R_1(Q_1)$ рангов, больших (меньших) s_1; справа от s_2 имеется $R_2(Q_2)$ рангов, больших (меньших) s_2, …, справа от s_{n-1} имеется $R_{n-1}(Q_{n-1})$ рангов, больших (меньших) s_{n-1}.

Введём обозначение суммы рангов:

$$R = R_1 + R_2 + \ldots + R_{n-1}; \;\; Q = Q_1 + Q_2 + \ldots + Q_{n-1}.$$

Выборочный коэффициент ранговой корреляции Кендалла находится по формуле:

$$\tau_B = \frac{4R}{n^2 - n} - 1 = \frac{2(R-Q)}{n^2 - n}. \qquad (4.6)$$

При $n \geq 10$ пользуются нормальным приближением для распределения τ: если $|\tau_B| \geq T_{\text{кр}} = z_{\text{кр}} \sqrt{\frac{2(2n+5)}{9n(n-1)}}$, то гипотеза независимости отклоняется, в противном случае принимается (здесь α – заданный уровень значимости, $z_{\text{кр}}$ – критическое значение порядка $\frac{\alpha}{2}$ стандартного нормального распределения).

Коэффициент конкордации (согласованности) Кендалла. Коэффициент конкордации Кендалла используется в случае, когда совокупность объектов характеризуется несколькими последовательностями рангов, а исследователю необходимо установить статистическую связь между этими последовательностями. Такие задачи возникают, например, при анализе экспертных оценок: несколько экспертов ранжируют одних и тех же испытуемых по определенному качеству, а исследователю для проведения углубленного анализа ситуации и принятия обоснованного решения требуется определить степень согласованности мнений группы экспертов.

Коэффициент конкордации Кендалла определяется по формуле

$$W = \frac{12 \cdot \sum_{i=1}^{n} D_i^2}{m^2 \cdot \left(n^3 - n\right)}. \qquad (4.7)$$

где n – число оцениваемы объектов (испытуемых), m – число ранговых последовательностей (количество экспертов), $D_i = d_i - \bar{d}$ – отклонение суммы рангов i-го объекта $d_i = \sum_{j=1}^{m} R_{ij}$ от средней суммы рангов всех объектов $\bar{d} = \frac{1}{n} \sum_{i=1}^{n} d_i$. Средняя сумма рангов всех объектов может быть вычислена по формуле $\bar{d} = \frac{m(n+1)}{2}$, которая используется для контроля.

Значения коэффициента конкордации заключены в интервале $0 \leq W \leq 1$.

При наличии одинаковых рангов у одного эксперта расчетная формула для коэффициента конкордации приобретает следующий вид:

$$W = \frac{12 \sum_{i=1}^{n} D_i^2}{m\left(n^3 - n\right) - m \sum_{j=1}^{m} T_j}, \quad T_j = \sum_{k=1}^{l} \left(t_k^3 - t_k\right).$$

В корректирующем члене для j-го эксперта через t_k обозначено число одинаковых значений в k-ой группе (связке), l – число связок (групп с одинаковыми значениями) в ранговой последовательности j-го эксперта.

Проверка нулевой гипотезы $H_0 : W = 0$ (мнения экспертов не согласуются друг с другом) при альтернативной $H_1 : W \neq 0$ (мнения экспертов согласуются) при $n \geq 7$ проводится с помощью критерия Пирсона «хи-квадрат». Эмпирическое значение $\chi^2 = m \cdot (n-1) \cdot W$ сравнивается с критическим $\chi^2(\alpha, n-1)$, вычисленным для числа степеней свободы $k = n-1$ и соответствующего уровня значимости α.

Критическая область критерия определяется равенством $\chi^2 \geq \chi^2(\alpha, n-1)$.

4.1.4. Нелинейная корреляция. Корреляционное отношение.

В тех случаях, когда корреляция между Y и X имеет явно выраженный нелинейный характер (об этом можно судить по форме диаграммы рассеивания) и объём выборки велик, данные наблюдения группируют и представляют их в виде корреляционной таблицы (таблица 4.3).

Здесь $x_1,\ldots,x_l; y_1,\ldots,y_m$ – значения признаков X и Y соответственно, а $n_{x1},\ldots,n_{xl}; n_{y1},\ldots,n_{ym}$ – соответствующие частоты, n_{ij} – частота, с которой встречается пара (x_i, y_j), $n = \sum_{i=1}^{l} \sum_{j=1}^{m} n_{ij}$.

Заполнение клеток корреляционной таблицы даёт довольно наглядное представление о характере зависимости между случайными величинами. Кроме того, при «ручных» расчётах сгруппированные данные заметно облегчают вычисление выборочных характеристик исследуемых случайных величин.

Таблица 4.3

X	Y						
	y_1	y_2	…	y_j	…	y_m	n_x
x_1	n_{11}	n_{12}	…	n_{1j}	…	n_{1m}	n_{x1}
x_2	n_{21}	n_{22}	…	n_{2j}	…	n_{2m}	n_{x2}
…	…	…	…	…	…	…	…
x_i	n_{i1}	n_{i2}	…	n_{ij}	…	n_{im}	n_{xi}
…	…	…	…	…	…	…	…
x_l	n_{l1}	n_{l2}	…	n_{lj}	…	n_{lm}	n_{xl}
n_y	n_{y1}	n_{y2}	…	n_{yj}	…	n_{ym}	n

При наличии компьютера корреляционная таблица составляется только в случае явно выраженной *нелинейной* зависимости, когда надо вычислить выборочные *корреляционные отношения* (эти характеристики могут быть найдены только по сгруппированным данным).

Корреляционное отношение определяется соотношением:

$$\eta = \frac{\sigma_{\text{М}}}{\sigma_{\text{О}}}. \qquad (4.8)$$

Если $\eta_{y/x}$ – корреляционное отношение случайной величины Y по случайной величине X, то:

$\sigma_M^2 = \dfrac{\sum\limits_{i=1}^{l} n_{xi}(\bar{y}_{xi} - \bar{y})^2}{n}$ – *межгрупповая* дисперсия, характеризует разброс условных средних $\bar{y}_{xi}$ от общей средней $\bar{y}$; $\sigma_{\text{О}}^2 = \sigma_y^2$ – *общая* дисперсия, характеризует разброс фактических данных y_j от их общей средней $\bar{y}$.

Если $\eta_{x/y}$ – корреляционное отношение случайной величины X по случайной величине Y, то:

$\sigma_M^2 = \dfrac{\sum\limits_{j=1}^{m} n_{yj}(\bar{x}_{yj} - \bar{x})^2}{n}$ – межгрупповая дисперсия, характеризует разброс условных средних $\bar{x}_{yj}$ от общей средней $\bar{x}$, $\sigma_{\text{О}}^2 = \sigma_x^2$ – общая дисперсия, характеризует разброс фактических данных x_i от их общей средней $\bar{x}$.

Корреляционное отношение обладает следующими свойствами.

1) $0 \le \eta_{y/x} \le 1;\ 0 \le \eta_{x/y} \le 1.$

2) Необходимое и достаточное условие отсутствия корреляционной зависимости в том, что $\eta_{y/x} = 0$.

3) Если $\eta_{y/x} = 1$, то между случайными величинами X и Y существует функциональная зависимость $y = f(x)$.

4) Коэффициент корреляции между величинами X и Y всегда по абсолютной величине не больше корреляционных отношений: $|r_B| \le \eta_{y/x};\ |r_B| \le \eta_{x/y}$.

ТЕМА 4.2 Регрессионный анализ

4.2.1 Парная линейная регрессия

Если корреляционный анализ позволяет оценит наличие и силу статистической взаимосвязи, то целью *регрессионного анализа* является установление формы этой зависимости. Такая форма определяется в виде некоторой функции зависимости величины Y от независимых величин $X_1, X_2, ..., X_k$ (факторов), которая называется *уравнением регрессии*.

Если исследуется зависимость случайной величины Y от одного фактора X, то модель называется *однофакторной* (или *парной*). Если же число независимых случайных величин два и больше $(k \geq 2)$, то регрессионная модель называется *многофакторной* или (*множественной*). Различают также *линейную* и *нелинейную* регрессию.

Линейная регрессионная модель.

Линейной регрессией называется сведение наблюдаемой на опыте зависимости некоторой переменной (*зависимой* или *объясняемой*) от одной или более других переменных (*независимых* или *объясняющих*) к линейной зависимости (в предположении, что строгая линейная зависимость между ними нарушается случайными ошибками). Для проведения линейной регрессии часто используется *метод наименьших квадратов*.

В простейшем случае речь идет о двух переменных. Пусть x – независимая переменная, y – зависимая и между ними существует следующая связь: $y_i = a_0 + a_1 x_i + \varepsilon_i$, где a_0 и a_1 – числовые коэффициенты, ε_i – случайные ошибки. При статистическом анализе линейной регрессионной модели предполагается также, что случайные ошибки наблюдений ε_i имеют нормальное распределение, т.е. $\varepsilon_i \sim N(0, \sigma);\ i = \overline{1,n}.$

В этом случае ошибки наблюдений ε_i также являются независимыми случайными величинами.

Задача состоит в том, чтобы по имеющимся наблюдениям $(x_1, y_1); (x_2, y_2); ...; (x_n, y_n)$ построить оценки для a_0 и a_1. Согласно методу наименьших квадратов, необходимо решить следующую математическую задачу: $S = \sum_{i=1}^{n} (y_i - (a_0 + a_1 x))^2 \to \min$.

Решаем задачу, вычисляя частные производные суммы квадратов по каждому из коэффициентов и приравнивая эти производные к нулю. Получаем систему *нормальных уравнений*, которая позволяет получить оценки параметров a_0 и a_1:

$$a_1 = \frac{K_{xy}}{s_x^2} = r_B \frac{s_y}{s_x};\quad a_0 = \bar{y} - a_1 \bar{x}. \tag{4.9}$$

Здесь:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i \; ; \;\; y_i = \frac{1}{n}\sum_{i=1}^{n} y_i \; ; \;\; K_{xy} = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})^2 \; ;$$

$$s_x^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2; \; s_y^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i - \bar{y})^2 \, .$$

Уравнение вида $\bar{y}_x = a_0 + a_1 x$ называется *уравнением линейной регрессии* Y на X, а получаемые из него значения $\widetilde{y}_i = a_0 + a_1 x_i$ называются *предсказанными* значениями, в отличие от *наблюдаемых* значений y_i.

Угловой коэффициент прямой линии регрессии Y на X (a_1) называют *выборочным коэффициентом регрессии* Y на X и обозначают $\rho_{y/x}$. В уравнениях линейной регрессии коэффициент $\rho_{y/x}$ характеризует чувствительность одного фактора при изменении другого фактора на одну единицу.

Замечание 4.1. Аналогично можно найти выборочное уравнение прямой линии регрессии X на Y: $\bar{x}_y = \rho_{x/y} y + c$, где $\rho_{x/y}$ – выборочный коэффициент регрессии X на Y.

В формуле (4.9) K_{xy} и r_B есть соответственно эмпирические *корреляционный момент (ковариация)* и *коэффициент корреляции* для величин X и Y.

Замечание 4.2. Так как коэффициенты линейной регрессии можно выразить через выборочный коэффициент корреляции r_B с помощью формул:

$$\rho_{y/x} = r_B \frac{s_y}{s_x}; \; \rho_{x/y} = r_B \frac{s_x}{s_y},$$

то уравнения линейной регрессии можно записать в виде:

$$\bar{y}_x - \bar{y} = r_B \frac{s_y}{s_x}(x - \bar{x}) \text{ и } \bar{x}_y - \bar{x} = r_B \frac{s_x}{s_y}(y - \bar{y}).$$

Качество аппроксимации (приближения) результатов наблюдений $(x_i; y_i)$ выборочной регрессии $\bar{y}_x = a_0 + a_1 x$ определяется величиной *остаточной дисперсии*, вычисляемой по формуле:

$$D_{\text{ост}} = \frac{1}{n-2}\sum_{i=1}^{n}(y_i - \bar{y}_x(x_i))^2 = \frac{Q_e}{n-2} \qquad (4.10)$$

Величина Q_e называется *остаточной суммой квадратов* (или *суммарной невязкой*). Разности между наблюдаемыми значениями переменной Y при $x = x_i$ и расчётными значениями $\bar{y}_x(x_i) = a_0 + a_1 x_i$ называют *остатками* и обозначают e_i.

Величина $\sigma_{\text{ост}} = \sqrt{D_{\text{ост}}}$ называется *стандартной ошибкой оценки по уравнению регрессии*. Стандартная ошибка оценки похожа на стандартное отклонение выборки, но не использует среднее значение. Чем меньше эмпирические данные рассеяны вокруг теоретических, тем меньше стандартная ошибка оценки. Эта ошибка характеризует влияние на величину результата неучтённых факторов.

Оценка существенности (значимости) уравнения регрессии в целом, т.е. проверка *адекватности* модели производится путем расчета *F*-критерия Фишера и сопоставления его с табличным (критическим). Для этого необходимо сравнить две суммы квадратов:

1) Остаточную сумму квадратов, характеризующую отклонение от регрессии Q_e.

2) Сумму квадратов, обусловленную регрессией $Q_R = \sum_{i=1}^{n} \left(\bar{y}_x(x_i) - \bar{y}\right)^2$.

Тогда выборочное значение *F*, имеющее распределение Фишера:

$$F = \frac{Q_R(n-2)}{Q_e} \tag{4.11}$$

Уравнение регрессии значимо, если $F_{\text{в}} > F_{\text{кр}}(\alpha, 1, n-2)$ с вероятностью $\gamma = 1 - \alpha$, где α – уровень значимости. В этом случае нулевой гипотезой H_0 является предположение о том, что уравнение регрессии не значимо. Следовательно, альтернативная гипотеза H_1 – уравнение регрессии значимо.

4.2.2. Нелинейная регрессия.

Во многих практических задачах зависимость между переменными *Y* и *x* нелинейна по параметрам. Однако часто можно найти преобразование переменных, которое приводит к линейной модели. Как правило, вычисление оценок параметров для линейной модели существенно упрощается.

Если экспериментальные точки располагаются вдоль некоторой линии, сходной по форме, например, с графиком гиперболической, показательной, логарифмической или других функций с неизвестными параметрами выбирается в качестве аппроксимирующей. Затем проводится *линеаризация* этой функции с помощью замены переменных, и задача сводится к аппроксимации зависимости многочлена первой степени.

Если выбирается структура аппроксимирующей функции в виде $\bar{y}_x = a_0 + a_1 \cdot \varphi(x)$, где $\varphi(x)$ – любая известная нелинейная функция, то можно сделать замену $t = \varphi(x)$ и рассматривать задачу линейного одномерного регрессионного анализа с аппроксимирующей функцией $\bar{y}_x = a_0 + a_1 \cdot t$.

Некоторые примеры таких замен:

1) $\bar{y}_x = a_0 + a_1\sqrt{x}$ – параболическая аппроксимация. Замена $t = \sqrt{x}$.

2) $\bar{y}_x = a_0 + \dfrac{a_1}{x}$ – гиперболическая аппроксимация. Замена $t = x^{-1}$.

3) $\bar{y}_x = a_0 + a_1 \ln x$ – логарифмическая аппроксимация. Замена $t = \ln x$.

4) $\bar{y}_x = a_0 + a_1 e^x$ – аппроксимация показательной функцией. Замена $t = e^x$.

Обычно рассматривают несколько видов функций $\bar{y}_x = \varphi(x, a_0, a_1)$ и выбирают ту функцию, для которой суммарная погрешность (невязка) окажется наименьшей.

Если аппроксимирующей функцией является квадратичная зависимость

$$\bar{y}_x = ax^2 + bx + c$$

то её параметры *a*, *b*, *c* находят из условия минимума функции:

$$S(a,b,c) = \sum_{i=1}^{n}\left(y_i - \left(ax_i^2 + bx_i + c\right)\right)^2 = \sum_{i=1}^{n}\left(y_i - ax_i^2 - bx_i - x\right)^2 .$$

Условия минимума функции сводятся к системе уравнений:

$$\begin{cases} \dfrac{\partial S}{\partial a} = 0 \\ \dfrac{\partial S}{\partial b} = 0 \\ \dfrac{\partial S}{\partial c} = 0 \end{cases} \Leftrightarrow \begin{cases} a\cdot\sum x_i^4 + b\cdot\sum x_i^3 + c\cdot\sum x_i^2 = \sum x_i^2 y_i \\ a\cdot\sum x_i^3 + b\cdot\sum x_i^2 + c\cdot\sum x_i = \sum x_i y_i \\ a\cdot\sum x_i^2 + b\cdot\sum x_i + c\cdot n = \sum y_i \end{cases},$$

при решении, которой находим искомые значения параметров *a, b* и *c*.

Эта система уравнений называется *нормальной системой способа наименьших квадратов при выравнивании по параболе*.

Можно ввести некоторый коэффициент связи, аналогичный корреляционному отношению, который характеризует силу связи между величиной *Y* и всеми аргументами аппроксимирующей функции в совокупности. Этот коэффициент называется коэффициентом детерминации.

$$R^2 = 1 - \frac{\sum_i (\tilde{y}_i - y_i)^2}{\sum_i (y_i - \bar{y})^2} \qquad (4.12)$$

где y_i – фактическое значение зависимой величины в i-м наблюдении; $\widetilde{y}_i$ – значение зависимой переменной, определяемой по уравнению регрессии; $\bar{y}$ – среднее арифметическое фактических значений зависимой переменной.

Для небольших значений $(n < 30)$ необходимо использовать скорректированный коэффициент детерминации $R^{*2} = 1 - \dfrac{n-1}{n-k-1}\left(1 - R^2\right)$.

Замечание 4.3. Величина R^2 является *оценкой корреляционного отношения* $\eta^2_{y/x}$. Если имеет место только линейная связь, то величина R^2 является оценкой квадрата коэффициента корреляции r^2_{XY}.

Коэффициент детерминации может принимать значения от 0 до 1. Чем больше коэффициент детерминации, тем точнее будет модель. В случае, когда $R^2 < 0{,}6$, считают, что точность приближения недостаточна и модель требует улучшения (введения новых независимых переменных, учета нелинейности и т.д.). Если коэффициент детерминации $R^2 \geq 0{,}9$, то регрессия считается достаточно точной для того, чтобы использовать ее для практических расчетов.

РАЗДЕЛ V. Непараметрическая статистика

ТЕМА 5.1 Непараметрические методы математической статистики

5.1.1 Критерии однородности, случайности и симметрии

В практике обработки результатов наблюдений распределение генеральной совокупности часто неизвестно либо (для непрерывных случайных величин) отличается от нормального распределения. В этих случаях применяют *методы не зависящие (или свободные) от распределения генеральной совокупности*, называемые также *непараметрическими методами*.

Непараметрические методы используют не численные значения элементов выборки, а *структурные свойства выборки* (например, отношения порядка между элементами). В связи с этим теряется часть информации, содержащаяся в выборке, поэтому, например, мощность непараметрических критериев меньше, чем мощность их аналогов, рассмотренных ранее. Но непараметрические методы могут применяться при более общих предположениях, и более просты с точки зрения выполнения вычислений.

Большая группа непараметрических критериев используется для проверки гипотезы о принадлежности двух выборок $x_1,\ldots,x_n$ и $y_1,\ldots,y_n$ одной и той же генеральной совокупности, то есть о том, что функции распределения $F_1(x)$ и $F_2(y)$ двух генеральных совокупностей равны: $F_1(x) \equiv F_2(y)\big|_{x=y}$.

Такие генеральные совокупности называются *однородными*. Необходимое условие однородности состоит в равенстве характеристик положения и

(или) рассеивания у рассматриваемых генеральных совокупностей – таких как средние, медианы, дисперсии и т.д. Используемые для этих целей непараметрические критерии в качестве основного предположения используют только непрерывность распределения генеральной совокупности.

Виды непараметрических статистических критериев при сравнении выборок приведены в таблице 5.1.

Таблица 5.1

Приложения	**Параметрический тест**	**Непараметрический тест**
Парные (зависимые выборки)	*t*-тест или *z*-тест	Критерий знаков Знако-ранговый Критерий (*T*-критерий Уилкоксона)
Две независимые выборки	*t*-тест или *z*-тест	Критерий Уилкоксона, Манна-Уитни

Критерий знаков применяется для проверки гипотезы H_0 об однородности генеральных совокупностей по попарно связанным (зависимым) выборкам.

Статистикой критерия знаков является число знаков ‘+’ или '–' в последовательности знаков разностей попарных выборок (x_i, y_i), $i = \overline{1,l}$ (l – число ненулевых разностей, $l < n$). В дальнейшем для определённости берётся число знаков ‘+’. При условии, что проверяемая гипотеза H_0 верная, число знаков ‘+’ имеет *биномиальное распределение* с параметрами $p = \frac{1}{2}$ и l. Задача сводится к проверке гипотезы $H_0 : p = \frac{1}{2}$ при одной из альтернативных гипотез:

$$H_1^{(1)} : p > \frac{1}{2};\ H_1^{(2)} : p < \frac{1}{2};\ H_1^{(3)} : p \neq \frac{1}{2}.$$

Часто более удобно проводить проверку гипотезы H_0, используя статистику Фишера.

Гипотеза H_0 отклоняется, если при $H_1^{(1)} : p > \frac{1}{2}$ выполняется неравенство:

$$F_B = \frac{r}{l - r + 1} \geq F(\alpha, k_1, k_2) \qquad (5.1)$$

где r – наблюдённое число знаков '+', – α –заданный уровень значимости, $k_1 = 2(l - r + 1);\ k_2 = 2r$.

Если $H_1^{(2)} : p < \frac{1}{2}$, то

$$F_B = \frac{l - r}{r + 1} \geq F(\alpha, k_1, k_2) \qquad (5.2)$$

где $k_1 = 2(r + 1);\ k_2 = 2(l - r)$.

При $H_1^{(3)} : p \neq \frac{1}{2}$ должно выполняться одно из неравенств (5.1) или (5.2) с заменой α на $\frac{\alpha}{2}$.

Знако-ранговый критерий (Т–*критерий Уилкоксона*). Этот критерий применяется для сравнения результатов, измеренных в двух разных условиях на одной и той же выборке (группе испытуемых): $x_1,\ldots,x_n$ и $y_1,\ldots,y_n$. Рекомендуется для выборок умеренной численности (численность каждой выборки от 12 до 40).

Вычисляют разности между индивидуальными значениями во втором и первом замерах («после» – «до»): $x_i - y_i$. Не обращая внимания на знак каждой разности, ранжируют их по порядку от наименьшей к наибольшей. В случае, когда две парные отметки совпадают, разность равна нулю. Эта нулевая разность не учитывается в присваивании рангов. Затем суммируются отдельно положительные и отрицательные ранги. Эмпирическое значение критерия равно меньшей по абсолютной величине сумме рангов. Далее эмпирическое (наблюдаемое) значение сравнивают с критическим, при этом в таблице при нахождении критического значения имеется в виду, что объем выборки равен числу только ненулевых разностей между отметками. При заданном уровне значимости α основная гипотеза отклоняется, если $|T_{\text{набл}}| \leq T_{\text{крит}}$.

Требования к *T*-критерию Уилкоксона более строгие, чем к критерию знаков. Однако если они удовлетворены, то критерий Уилкоксона имеет большую мощность, чем критерий знаков.

Критерий Уилкоксона, Манна-Уитни (*ранговый U-критерий*). Он был разработан в 1945 году Ф. Уилкоксоном, а в 1947 году существенно переработан и расширен Х. Б. Манном и Д. Р. Уитни.

Критерий проверяет гипотезу о том, что выборки извлечены из общей генеральной совокупности. В частности, он применим для проверки гипотезы о равенстве средних для *независимых* выборок.

Предполагается, что объем первой выборки меньше (не больше) объёма второй: $n_1 \le n_2$; если это не так, то выборки можно перенумеровать (поменять местами).

Проверка нулевой гипотезы в случае, если объём обеих выборок не превосходит 25.

1. Для того чтобы при заданном уровне значимости $\alpha = 2Q$ проверить нулевую гипотезу $H_0 : F_1(x) = F_2(x)$ об однородности двух независимых выборок объёмов n_1 и n_2 при конкурирующей гипотезе $H_1 : F_1(x) \ne F_2(x)$ надо:

1) расположить варианты обеих выборок в возрастающем порядке, т. е. в виде одного вариационного ряда, и найти в этом ряду наблюдаемое значение критерия W_H — сумму порядковых номеров вариант первой выборки;

2) найти по таблице приложения (см. [4]) нижнюю критическую точку $w_{\text{нижн.кр}}(Q, n_1, n_2)$, где $Q = \dfrac{\alpha}{2}$;

3) найти верхнюю критическую точку по формуле $w_{\text{верхн.кр.}} = (n_1 + n_2 + 1) \cdot n_1 - w_{\text{нижн.кр}}$.

Если $W_H < w_{\text{нижн.кр}}$ или $W_H > w_{\text{верхн.кр}}$ — нулевую гипотезу отвергают.

Если $w_{\text{нижн.кр}} < W_H < w_{\text{верхн.кр.}}$ — нет оснований отвергнуть нулевую гипотезу.

2. При конкурирующей гипотезе $H_1 : F_1(x) > F_2(x)$ надо найти по таблице критическую точку $w_{\text{нижн.кр}}(Q, n_1, n_2)$, где $Q = \alpha$. Если $W_H > w_{\text{нижн.кр.}}$ – нет оснований отвергнуть нулевую гипотезу. Если $W_H < w_{\text{нижн.кр.}}$ – нулевую гипотезу отвергают.

3. При конкурирующей гипотезе $H_1 : F_1(x) < F_2(x)$ надо найти по таблице критическую точку: $w_{\text{верхн.кр}}(Q, n_1, n_2) = (n_1 + n_2 + 1) \cdot n_1 - w_{\text{нижн.кр}}(Q, n_1, n_2)$, где $Q = \alpha$. Если $W_H < w_{\text{верхн.кр.}}$ – нет оснований отвергнуть нулевую гипотезу. Если $W_H > w_{\text{верхн.кр.}}$ – нулевую гипотезу отвергают.

Принятие конкурирующей гипотезы $H_1 : F_1(x) > F_2(x)$ означает, что $X < Y$. Аналогично, если справедлива гипотеза $H_1 : F_1(x) < F_2(x)$, то $X > Y$.

Замечание 5.1. При вычислении наблюдаемого значения критерия Уилкоксона следует учесть, что ранги совпадающих вариант *различных* выборок равны среднему арифметическому порядковых номеров вариант в общем вариационном ряде, составленном из вариант обоих выборок.

Замечание 5.2. Расчётное значение критерия Манна-Уитни определяется в соответствии со следующей формулой:

$$U = n_1 \cdot n_2 + \frac{n_x \cdot (n_x + 1)}{2} - S_x, \qquad (5.3)$$

где S_x – наибольшая из ранговых сумм, n_x – объём выборки, которой соответствует эта сумма. По таблице критических точек находится критическое значение (по данным n_1 и n_2). Если полученное значение меньше табличного или равно ему для выбранного уровня значимости, то принимается альтернативная гипотеза.

Замечание 5.3. Статистики U и W связаны равенством:

$$W = U + \frac{n_1 \cdot (n_1 + 1)}{2}.$$

5.1.2. Непараметрические критерии. Факторный анализ

В дисперсионном анализе используется критерий Фишера, чтобы сравнивать средние трех и более совокупностей; предполагается, что совокупности нормально распределены и что дисперсии совокупностей равны. Когда эти условия не выполняются, то для сравнения трех и более средних может использоваться *непараметрический критерий Краскела–Уоллиса*. Для применения критерия требуется соблюдение *двух условий*:

1) Выборки независимы и получены случайным образом.

2) Размер каждой выборки должен быть не меньше пяти.

В этом случае исследуемое распределение приближается к хи-квадрат распределению с числом степеней свободы $k = m - 1$, где m – число уровней исследуемого признака (столбцов таблицы). Для выборок меньшего размера требуются специальные таблицы.

При этом не требуется, чтобы генеральные совокупности имели нормальный закон распределения.

Суть метода состоит в следующем. В критерии Краскела–Уоллиса все выборки перемешиваются и их значения ранжируются. Далее вычисляются средние ранги для каждой выборки и средний ранг по всем данным. Для проверки используется следующая статистика (два варианта формулы):

$$H = \frac{12}{N \cdot (N+1)} \sum_{i=1}^{m} \frac{R_i^2}{n_i} - 3 \cdot (N+1) \text{ или } H = \frac{12}{N \cdot (N+1)} \sum_{i=1}^{m} n_i \left(\widetilde{R}_i - \overline{\overline{R}} \right)^2,$$

где R_i – сумма рангов по столбцу, $\widetilde{R}_i$ – средний ранг по столбцу, $\overline{\overline{R}}$ – средний ранг по выборке, $\overline{\overline{R}} = \frac{N+1}{2}$; $N = n_1 + n_2 + \ldots + n_m$ – суммарный объем всех m выборок; n_i – количество измерений на уровне i.

Если выборки взяты из различных совокупностей, их средние ранги могут сильно различаться, значение H будет велико – нулевую гипотезу следует отвергнуть. Для двух выборок критерий совпадает с уже известным критерием Уилкоксона.

Если в объединенном вариационном ряду имеются *связки* (группы, состоящие из совпавших величин), то статистику H нужно заменить статистикой $H^* = \frac{H}{a}$, где $a = 1 - \frac{1}{(N+1)\cdot\left(N^2-1\right)}\sum_{s=1}^{l} t_s \cdot \left(t_s^2 - 1\right)$. Здесь:

- l – число связок;
- t_s – число наблюдений в s-связке.

Критерий Фридмана используется в тех случаях, когда не выполняются предположения, на которых основан классический дисперсионный двухфакторный анализ.

Для применения критерия Фридмана необходимо выполнять следующие условия:

- Измерение должно быть проведено в шкале интервалов или отношений;
- Выборка должно быть связной (зависимой);
- В выборке должно быть не менее двух испытуемых, каждый из которых имеет не менее трех измеренных показателей. Верхний предел для количества испытуемых не определен, а количество измерений не может превышать 100;
- В зависимости от числа измерений и количества испытуемых используются разные таблицы значимости.

Пусть таблица результатов оценки или наблюдений n объектов состоит из n строк и m столбцов. В строках записываются m ранжированных переменных, причем длины ранжировок (объемы выборки) равны n. Строки таблицы можно рассматривать как m связанных выборок объемом n. Связность выборок следует из того, что выборки – суть повторные наблюдения на одних и тех же n объектах. Если объекты не различаются между собой, суммы рангов по столбцам также не будут различаться. Нулевая гипотеза H_0: между столбцами нет различия – проверяется с помощью статистики Фридмана S.

Выборочное значение S-статистики S_B вычисляется по формуле:

$$S_B = \frac{12}{mn(n+1)}\sum_{j=1}^{n}\left[\sum_{i=1}^{m} r_{ij} - \frac{1}{2}m(n+1)\right]^2 = \frac{12}{mn(n+1)}\sum_{j=1}^{n}\left(\sum_{i=1}^{m} r_{ij}\right)^2 - 3m(n+1),$$

где r_{ij} – ранг j-го объекта, присваиваемый i-м экспертом.

Если гипотеза H_0 верна, то при $m \to \infty$ статистика S имеет распределение хи-квадрат с $(n-1)$ степенями свободы.

Гипотеза H_0 отклоняется на уровне значимости α, если $S_B > \chi_\alpha^2(n-1)$ – квантиль распределения хи-квадрат порядка α.

Мерой согласия различных ранжировкок n объектов является коэффициент конкордации (согласия) Кендалла $W = \dfrac{S}{m(n-1)}$ (см. раздел IV, тема 4.1).

В случае, когда в ранжировках (в строках таблицы) имеются совпадающие ранги, вычисляется скорректированная статистика:

$$S' = \frac{\sum_{j=1}^{n}\left[\sum_{i=1}^{m} r_{ij} - \dfrac{m(n+1)}{2}\right]}{\dfrac{mn(n+1)}{12} - \dfrac{1}{n-1}\sum_{i=1}^{m} T_i},$$

где $T_i = \dfrac{1}{12}\sum_{t=1}^{l}\left(n_t^3 - n_t\right), i = \overline{1,m}$. Здесь l – число групп повторяющихся рангов в i-й ранжировке; n_t – число совпадающих рангов в группе с номером $t = \overline{1,l}$.

РАЗДЕЛ VI Задачи прогнозирования

ТЕМА 6.1 Временные ряды и множественная линейная регрессия

6.1.1 Временные ряды.

Основные понятия. Рассмотрим случайный объект или явление, характеристика Y которого меняется во времени. Выполнив n наблюдений над этим объектом в равноотстоящие моменты времени $t_1, t_2, \ldots, t_n$, получим упорядоченную последовательность чисел $y(t_1), y(t_2), \ldots, y(t_n)$, которая называется *временным (динамическим) рядом* или *случайной последовательностью.* Чаще используется более компактная форма временных рядов $y_1, y_2, \ldots, y_n$ (здесь $y_i \equiv y(t_i), i = \overline{1,n}$). Числовые значения y_i при этом называются *уровнями ряда*.

Примерами временных рядов могут служить:

- результаты ежесуточных замеров солености воды в определенной точке мирового океана;
- данные о среднесуточной температуре воздуха в конкретном населенном пункте;
- данные о курсе доллара на ММВБ;
- данные о числе сообщений, переданных за сутки в определенном направлении связи и т.д.

Временной ряд имеет два существенных отличия от простой выборки:

1. Элементы $x_1, x_2, \ldots, x_n$ случайной выборки взаимно независимы, тогда как значение y_i временного ряда, зафиксированное в момент t_i, может

существенно зависеть от одного или нескольких значений ряда $y_1, y_2, \ldots, y_{i-1}$, зафиксированных до этого момента.

2. Элементы $x_1, x_2, \ldots, x_n$ случайной выборки имеют один и тот же закон распределения, между тем закон распределения i - го члена временного ряда (случайной величины y_i) может изменяться при изменении его номера i.

Уровни временного ряда могут характеризовать значение показателя на определенный момент времени (*моментные ряды*): например, температура воздуха, измеренная ежедневно в 12 часов дня. Если каждое значение уровня ряда образуется как сумма или среднее значение показателя за некоторый интервал времени, то такие ряды называются *интервальными*: например, временные ряды, отражающие значения среднемесячной заработной платы рабочих предприятия.

В структуре временного ряда обычно выделяют четыре основных элемента:

- тренд;
- сезонность;
- цикличность;
- случайная остаточная компонента (шум).

Любой ряд можно описать в виде комбинации всех или нескольких этих элементов.

Трендом называют устойчивое систематическое изменение показателя. С математической точки зрения, тренд описывается достаточно гладкой функцией от времени.

Сезонность – это систематически повторяющиеся колебания показателя, обусловленные временем года.

Цикличность – это регулярные колебания относительно тренда, обусловленные некоторыми постоянно действующими факторами. Эти колебания могут быть предсказаны и не связаны с временем года.

Случайная остаточная компонента обусловлена действием случайных факторов, влияющих на показатель. Она затрудняет обнаружение в структуре ряда регулярных компонент.

Методы сглаживания временных рядов. Методы сглаживания позволяют уменьшить влияние случайной компоненты временного ряда и таким образом выявить тренд и другие регулярные компоненты. Суть различных способов сглаживания сводится к замене фактических значений ряда y_i ряда расчетными значениями $\tilde{y}_i$, имеющими значительно меньшие колебания, чем исходные фактические значения. В ряде случаев сглаживание временного ряда является важным вспомогательным средством, заметно облегчающим применение других методов анализа этих рядов (в частности, аналитических методов выделения тренда). Рассмотрим два метода сглаживания: *метод скользящих средних* и *метод экспоненциального сглаживания*.

В *методе скользящих средних* каждый член ряда заменяется средним m соседних членов, т.е. рассчитывается по формуле:

$$\bar{y}_t = \frac{y_{t-m+1} + y_{t-m+2} + \ldots + y_t}{m},$$

где m – интервал сглаживания. При этом первые $m-1$ значений сглаженного ряда не рассчитываются.

Чаще всего сглаживание проводят по 3, 5 или 7 членам исходного ряда (нечетный интервал сглаживания). Чем больше интервал сглаживания, тем сильнее усреднение данных и тем больше учитываются предыдущие значения исследуемого показателя.

Метод экспоненциального сглаживания позволяет при расчете очередного сглаженного значения учесть всю «предысторию» развития данного показателя. При этом учитывается степень старения данных: чем старее информация, тем с меньшим весом входит она в формулу для расчета сглаженного значения. Сглаженное значение ряда (экспоненциальная средняя) рассчитывается по формулам:

$$Q_1 = y_1;\ Q_t = \alpha \cdot y_t + (1-\alpha) \cdot Q_{t-1}\ \left(t = \overline{2,n}\right),$$

где Q_t – экспоненциальная средняя в момент времени t; y_t – фактическое значение показателя в момент t; Q_{t-1} – предыдущее значение экспоненциальной средней; α – параметр сглаживания, характеризующий вес текущего (самого нового) наблюдения $(0 \le \alpha \le 1)$.

Если $\alpha = 1$, то предыдущие наблюдения полностью игнорируется, а если $\alpha = 0$, то игнорируется текущее наблюдение. Обычно используется α в диапазоне от 0,1 до 0,3. При выборе α необходимо учитывать, что для того чтобы сглаженный ряд прошел ближе к фактическим данным, нужно повысить значение α (тем самым увеличивается вес текущих наблюдений). Но при этом ряд становится менее гладким.

При использовании метода экспоненциального сглаживания возникает проблема определения начального сглаженного значения $\tilde{y}_0$. В качестве начального сглаженного значения обычно используют первый член исследуемого временного ряда, т.е. $\tilde{y}_0 = y_1$.

Аналитическое сглаживание временных рядов. Модели тренда. Метод скользящего среднего и метод экспоненциального сглаживания облегчают выявление тренда исследуемого временного ряда. Но ряд сглаженных значений громоздок (он содержит практически столько значений, сколько и сам исходный временной ряд) и не может быть использован при аналитическом решении задач анализа временных рядов. Поэтому возникает необходимость «компактного» описания тренда при помощи некоторой функции времени, в функциональной форме и параметрах которой концентрировалась бы вся су-

щественная информация о тенденции развития временного ряда. Такая функция называется *математической моделью тренда*. Процесс подбора математической модели тренда по данным наблюдения называют *аналитическим сглаживанием временного ряда*. Аналитическое сглаживание временного ряда выполняется в следующем порядке: сначала выбирается тип сглаживающей функции, затем определяются выборочные оценки параметров этой функции.

Первым шагом в построении функции тренда обычно является сглаживание временного ряда. Затем по виду полученного графика выбирают одну или несколько моделей, называемых *функциями-кандидатами*.

Простейшими математическими моделями тренда, широко используемыми при анализе временных рядов, являются следующие модели:

- *Линейная* $\tilde{y}(t) = a_0 + a_1 t$. Эта модель описывает тренд, скорость изменения которого постоянна и равна a_1. При $a_1 > 0$ тренд равномерно возрастает, при $a_1 < 0$ $a_1 < 0$ – равномерно убывает.
- *Логарифмическая* $\tilde{y}(t) = a_0 + a_1 \ln t$, описывающая тренд с постепенным уменьшением скорости роста.
- *Полиномиальная* $\tilde{y}(t) = a_0 + a_1 t + a_2 t^2 + \ldots + a_m t^m$, где m – степень полинома. Частный случай этой модели – полином второй степени $\tilde{y}(t) = a_0 + a_1 t + a_2 t^2$ описывает тренд с постоянным ускорением изменения, равным $2a_2$. При $a_2 > 0$ скорость изменения тренда возрастает, при $a_2 < 0$ – убывает. Полином третьей степени $\tilde{y}(t) = a_0 + a_1 t + a_2 t^2 + a_3 t^3$ описывает тренд с переменным изменением ускорения. При $a_3 > 0$ ускорение возрастает, при $a_3 < 0$ – убывает.
- *Степенная* $\tilde{y}(t) = a_0 \cdot t^{a_1}$.
- *Экспоненциальная* $\tilde{y}(t) = a_0 \cdot e^{a_1}$, описывающая тренд, у которого скорость и ускорение изменения пропорциональны величине самого тренда.

Наиболее распространенным методом получения «наилучших» оценок неизвестных параметров сглаживающих функций является *метод наименьших квадратов*.

После определения параметров функций-кандидатов оценивается точность моделей как совокупная разница между фактическими значениями показателя и его соответствующими теоретическими значениями. В качестве показателя точности трендовой модели может использоваться сумма квадратов отклонений $\sum_{i=1}^{n} (y_i - \tilde{y}(t_i))^2$, которая была минимизирована при расчете параметров тренда.

Но чаще точность оценивается на основании *коэффициента детерминации* R^2: $R^2 = 1 - \dfrac{\sum_{i=1}^{n}(y_i - \tilde{y}(t_i))^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$, где n – количество уровней временного ряда (число наблюдений); y_i – фактическое значение показателя в момент времени t_i; $\tilde{y}(t_i)$ – теоретическое (рассчитанное по тренду) значение показателя в момент времени t_i; $\bar{y}$ – среднее арифметическое фактических значений, которое рассчитывается по формуле $\bar{y} = \dfrac{1}{n}\sum_{i=1}^{n} y_i$.

Коэффициент детерминации всегда удовлетворяет условию $0 \le R^2 \le 1$. Чем больше R^2 (ближе к единице), тем точнее модель.

Среди функций-кандидатов выбирают наиболее точную модель (с наибольшим коэффициентом детерминации). Именно эту модель используют в дальнейшем для выполнения прогнозов.

Прогнозы, получаемые на основе трендовых моделей, можно разделить на точечные и интервальные.

Точечный прогноз дает единственное значение прогнозируемого показателя. Он получается подстановкой в уравнение выбранного тренда значения времени, относящегося к будущему.

Интервальный прогноз для каждого момента времени дает некоторый интервал значений, в котором можно ожидать появление прогнозируемой величины с заданной вероятностью. Этот прогноз осуществляется путем расчета доверительных интервалов.

6.1.2. Множественная регрессия.

Множественная корреляция является одним из немногих количественных методов, которые могут быть использованы для исследования взаимосвязей природных процессов, в том числе для оценки одновременного влияния нескольких факторов на данный процесс с целью его прогнозов и расчётов. Кроме того, этот метод позволяет определять относительное влияние на прогноз каждого фактора и измерять полный эффект с помощью коэффициентов. Можно также оценить значимость связи между зависимой и каждой независимой переменной и получить «лучшее» расчётное уравнение.

Модель множественной линейной регрессии – это уравнение вида

$$Y = a_0 + a_1 X_1 + a_2 X_2 + \ldots + a_m X_m,$$

где Y – *зависимая переменная* (предиктант, отклик); a_0 – *константа*; X_i – независимые переменные (предикторы, факторы); $i = \overline{1,m}$, m – количество предикторов.

Слово «предиктор» произошло от англ. *predict* – предсказывать. Процедуры множественной регрессии будут оценивать (вычислять) *параметры уравнения*, то есть коэффициенты $a_0, a_1, \ldots, a_m$. Величины $a_1, a_2, \ldots, a_m$ называются также *регрессионными коэффициентами*.

Требования к рядам наблюдений. Для получения удовлетворительных результатов при использовании модели множественной регрессии необходимо выполнение ряда требований к исходной информации, соблюдение которых зачастую вообще не проверяется, в то время как во многих случаях они не выполняются или выполняются не полностью.

Основные требования к рядам наблюдений заключаются в следующем:

1. Корреляция между прогнозируемым рядом Y (предиктантом) и каждым из независимых переменных X_i (предикторами) должна быть высокой – не менее 0.7.

2. Корреляция между рядами-предикторами, наоборот, должна отсутствовать или быть незначительной. При наличии тесной связи между предикторами корреляционная матрица становится вырождающейся, её определитель стремится к нулю, и возникают трудности в вычислении коэффициентов уравнения регрессии. В этом случае надо исключать дублирующие предикторы.

3. Связи между всеми рядами должны быть линейными. Если нелинейность связи очевидна, то можно рассмотреть или преобразования переменных, или явно допустить включение нелинейных членов.

4. Сопоставляемые ряды должны подчиняться нормальному закону распределения. Близость законов распределения выборок к нормальному является одним из главных показателей надёжности математических моделей, основанных на принципе метода наименьших квадратов.

5. Ряд-предиктант должен представлять собой выборку значений случайной величины, т.е. его значения должны быть некоррелированы между собой.

6. Объём выборки должен в несколько раз превосходить число независимых переменных. Практика показывает, что при использовании одного предиктора длина рядов n должна быть не менее 10, при двух предикторах минимальная длина рядов должна составлять не менее 25–30, при четырёх – 50–60, при пяти – 100–120 и т.д. Только в этом случае можно получить более или менее надёжные оценки параметров уравнения регрессии.

Предсказанные значения и остатки. Линия регрессии выражает наилучшее предсказание зависимой переменной Y по независимым переменным X_i. Однако природа редко бывает предсказуемой и обычно имеется существенный разброс наблюдаемых точек относительно подогнанной пря-

мой. Отклонение отдельной точки от линии регрессии (от предсказанного значения) называется *остатком*.

Анализ остатков является одним из способов проверки качества модели или степени ее адекватности данным. Если остатки представляют собой временной ряд случайных независимых величин, распределенных по нормальному закону, то это может служить обоснованием пригодности уравнения для прогноза. На графике остатки должны вести себя достаточно хаотично, не должно быть резких выбросов, закономерностей в чередовании знаков.

В частности, *выбросы* в данных (т.е. экстремальные наблюдения) могут вызвать серьезные ошибки в вычислении коэффициентов уравнения, «сдвигая» линию регрессии в определенном направлении. Часто исключение всего одного экстремального наблюдения приводит к совершенно другому результату.

Наличие на графике ряда остатков тренда или периодичности является признаком того, что в уравнении регрессии не учтены какие-то факторы, существенные для формирования данного процесса.

Чем меньше разброс значений остатков около линии регрессии по отношению к общему разбросу значений, тем лучше прогноз.

В случае множественной линейной корреляции важную роль играет величина R^2 – *коэффициент детерминации (определенности)* как показатель качества модели или применимости данного набора предикторов для описания зависимой переменной Y.

Значение R^2 является индикатором степени подгонки модели к данным. Коэффициент детерминации непосредственно интерпретируется следующим образом. Если $R^2 = 0{,}4$, то только 40% от исходной изменчивости ряда Y могут быть объяснены предикторами (X_i), а 60% остаются необъясненными. Таким образом, величина R^2 есть доля дисперсии исследуемой переменной Y, объяснённая переменными X_i.

Необъясненная (остаточная) доля дисперсии – это результат влияния или других параметров, не учтенных в модели, или между переменными существуют сложные нелинейные взаимосвязи.

В идеале желательно иметь объяснение если не для всей, то хотя бы для большей части исходной изменчивости. В статистике принято считать, что уравнение регрессии с данным набором предикторов можно использовать, если предикторы обеспечивают хотя бы 50% исходной дисперсии, т.е. при $R^2 \geq 0{,}5$. Значение R^2, близкое к 1, показывает, что модель объясняет почти всю изменчивость соответствующей переменной.

Интерпретация коэффициента множественной корреляции R. Обычно, степень зависимости двух или более предикторов X_i с зависимой переменной Y выражается с помощью *коэффициента множественной корреляции* R. Коэффициент множественной корреляции R можно интерпретировать как

парный коэффициент корреляции между двумя рядами Y: наблюдаемыми и вычисленными по уравнению регрессии. Это неотрицательная величина, принимающая значения между 0 и 1.

Если при добавлении еще одного предиктора коэффициент множественной корреляции R уменьшился, значит, этот предиктор ухудшает точность уравнения регрессии и его надо исключить из набора предикторов.

Для интерпретации направления связи между переменными смотрят на знаки регрессионных коэффициентов (или B-коэффициентов). Если B-коэффициент положителен, то связь этой переменной с зависимой прямая; если B-коэффициент отрицателен, то связь обратная. Если B-коэффициент равен нулю, связь между переменными отсутствует.

II ПРАКТИЧЕСКИЙ РАЗДЕЛ
2.1. ЛАБОРАТОРНЫЕ РАБОТЫ

Лабораторная работа № 1. Числовые характеристики случайных величин. Вероятностные распределения. Знакомство с пакетом Statistica.

Необходимые теоретические сведения для выполнения лабораторной работы находятся в теоретическом разделе – тема 1.1.

Контрольный пример 1.1.

Число α-частиц, достигающих счетчика, является случайной величиной X, распределенной по закону:

X	0	1	2	3	4	5	6	7	8	9	10
p	0,021	0,081	0,156	0,201	0,195	0,151	0,097	0,054	0,026	0,011	0,007

Пользуясь пакетом MS Excel, найти числовые характеристики случайной величины X: математическое ожидание $M(X)$, дисперсию $D(X)$, среднее квадратическое отклонение $\sigma(X)$, коэффициент асимметрии $As(X)$ и эксцесс $Ek(X)$; указать их размерность; построить многоугольник распределения случайной величины.

Решение.

1. Создадим массивы данных в MS Excel.

Это – массив значений случайной величины (**A2:A12**) и массив вероятностей (**B2:B12**), с которыми случайная величина X принимает свои значения (рис. 1.1).

	A	B	C
1	***X***	***p***	
2	0	0,021	
3	1	0,081	
4	2	0,156	
5	3	0,201	
6	4	0,195	
7	5	0,151	
8	6	0,097	
9	7	0,054	
10	8	0,026	
11	9	0,011	
12	10	0,007	
13	**Сумма**	**1**	
14	**M(X)**		
15	**D(X)**		

Рис. 1.1 – Исходные данные

2. Найдем математическое ожидание случайной величины *X*:

В MS Excel для этого можно использовать функцию **СУММПРОИЗВ (массив 1, массив 2)**, которая позволяет найти сумму произведений диапазонов ячеек:

1) Выделим ячейку **B14,** в которую будет возвращено значение функции.

2) На вкладке «Главная» нажатием кнопки «f_x» откроем диалоговое окно мастера функций и в категории «*Математические*» выберем функцию **СУММПРОИЗВ (A2:A12, B2:B12)**.

3) После нажатия кнопки *OK* в ячейке **B14** получаем значения математического ожидания случайной величины (рис. 1.2).

3. Найдем дисперсию *D(X)* случайной величины *X*:

$$D(X)=\sum_{i=1}^{n}(x_i-M(X))^2\cdot p_i$$

Выделим ячейку **C2**, в которой создадим формулу: =A2-$B$14. После нажатия клавиши «*Enter*» в ячейке **C2** появится число –3,868.

Скопируем формулу протягиванием ячейки **C2** указателем мыши за правый нижний угол табличного курсора (при нажатой левой кнопки мыши) до ячейки **C12** (см. рис 1.2).

Выделим ячейку **B15**, в которую будет возвращено значение дисперсии. В выделенную ячейку вводим функцию **СУММПРОИЗВ (C2:C12, C2:C12, B2:B12)**.

После нажатия кнопки OK в ячейке **B15** получаем значение дисперсии случайной величины (см. рис. 1.2): $D(X)=3{,}8406\approx 3{,}8$ (частиц^2).

4. Вычислим среднее квадратическое отклонение случайной величины:

$\sigma=\sqrt{D(X)}=\sqrt{3{,}8406}\approx 1{,}96\approx 2{,}0$ (частиц)

5. Найдем коэффициент асимметрии:

$$As(X)=\frac{1}{\sigma^3}\sum_i p_i(x_i-M(X))^3 .$$

B15

	A	B	C
1	***X***	***p***	***X-M(X)***
2	0	0,021	-3,868
3	1	0,081	-2,868
4	2	0,156	-1,868
5	3	0,201	-0,868
6	4	0,195	0,132
7	5	0,151	1,132
8	6	0,097	2,132
9	7	0,054	3,132
10	8	0,026	4,132
11	9	0,011	5,132
12	10	0,007	6,132
13	**Сумма**	**1**	
14	**M(X)**	**3,868**	
15	**D(X)**	**3,8406**	

Рис. 1.2 – вычисление *M(X)* и *D(X)*

Выделим ячейку **D2**, в которой создадим формулу: **=C2^3*B2.**

После нажатия клавиши «*Enter*» в ячейке **D2** появится число – 1,215.

Скопируем полученную формулу в ячейки **D3:D12**.

С помощью автосуммирования (кнопка «Σ» на вкладке «Главная») найдем сумму ячеек столбца **D**.

Результат (3,479) появляется в ячейке **D13** (см. рис. 1.3).

Вычислим коэффициент асимметрии:

$$As(X) \approx \frac{1}{1{,}96^3} \cdot 3{,}479 \approx 0{,}462 \approx 0{,}5.$$

6. Аналогичным образом вычислим эксцесс случайной величины (см. рис. 1.3): $Ek(X) = \frac{1}{\sigma^4} \sum_i p_i \left(x_i - M(X)\right)^4 - 3 \approx \frac{44{,}749}{1{,}96^4} - 3 \approx 0{,}032.$

E15 =E13/1,96^4-3

	A	B	C	D	E
1	***X***	***p***	***X-M(X)***	***p*(X-M(X))^3***	***p*(X-M(X))^4***
2	0	0,021	-3,868	-1,215	4,701
3	1	0,081	-2,868	-1,911	5,480
4	2	0,156	-1,868	-1,017	1,899
5	3	0,201	-0,868	-0,131	0,114
6	4	0,195	0,132	0,000	0,000
7	5	0,151	1,132	0,219	0,248
8	6	0,097	2,132	0,940	2,004
9	7	0,054	3,132	1,659	5,196
10	8	0,026	4,132	1,834	7,579
11	9	0,011	5,132	1,487	7,630
12	10	0,007	6,132	1,614	9,897
13	**Сумма**	**1**		**3,479**	**44,749**
14	**M(X)**	**3,868**		**As**	**Ex**
15	**D(X)**	**3,8406**		**0,462**	**0,032**

Рис. 1.3 – вычисление коэффициентов асимметрии и эксцесса

7. Построим многоугольник распределения случайной величины (рис. 1.4). Для этого выделим ячейки со значениями вероятностей **B2:B12**. На вкладке «*Вставка*» в группе *«График»* (для EXCEL 2016 – «*Диаграммы*») выберем тип графика – *график с маркерами*.

Щелкнем правой кнопкой мыши на поле графика и выберем *«Выбрать данные»*. В появившемся окне «Выбор источника данных», справа, где *«Подписи горизонтальной оси (категории)»*, выбираем *«Изменить»*.

В открывшемся окне «Диапазон подписей оси» добавляем массив значений случайной величины **A2:A12** и нажимаем *ОК*.

На вкладке «*Макет*» в разделе «*Названия осей*» выбираем подписи осей: *x* и *p*. На той же вкладке выбираем «*Название диаграммы*» и указываем название графика: «*Многоугольник распределения случайной величины X*» (рис. 1.4).

Нужный макет графика можно выбрать также на вкладке «*Конструктор*». Редактирование графика производится после щелчка правой кнопкой мыши на нужном элементе диаграммы.

Рис. 1.4 – Многоугольник распределения исходной случайной величины

Контрольный пример 1.2. В пакете MS Excel необходимо:

1. Ввести в таблицу значения аргумента *x* в диапазоне от –3 до 5 с шагом 0,2.

2. Вычислить значение плотности стандартного нормального распределения, а также плотности нормального распределения с параметрами $a = 2$, $\sigma = 1$; $a = 0$, $\sigma = 0{,}5$; $a = 1$, $\sigma = 2$.

3. Для заданных параметров нормального распределения построить семейство графиков плотности и функции распределения.

Решение. Для вычисления значений плотности и функции нормального распределения в Excel используется встроенная статистическая функция НОРМ.РАСП.

Для построения графика плотности распределения протабулируем функцию $f(x)$ на отрезке (–3; 5) с шагом 0,2. Для этого вводим в **A3** подпись «*x*», в ячейки **B3:E3** подписи, соответствующие условию задачи (рис.1.6).

Вводим в **A4** значение –3, в **A5** – значение –2,8, обводим, выделяя, ячейки **A4** и **A5** и, захватив нижний правый угол рамки вокруг ячеек **A4** и **A5**, перетягиваем его вниз до ячейки **A44**, что позволит автоматически занести в столбец значения от –3 до 5 с шагом 0,2 (на рис.1.6 приведена часть значений отрезка).

Ставим курсор в ячейку B4 и вызываем функцию плотности нормального распределения. Для этого нажимаем кнопку мастера функций f_x, выбираем категорию «*Статистические*» и функцию НОРМ.РАСП.

Появляется диалоговое окно функции. Заполняем ее как показано на рис. 1.5.

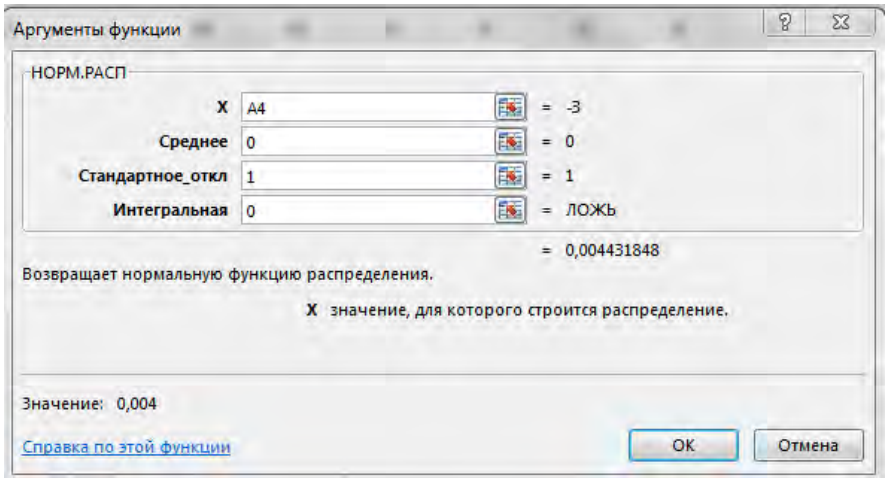


Рис. 1.5 – Диалоговое окно функции НОРМ.РАСП для *N*(0,1)

В поле «*Интегральная*» ставим ноль, нажимаем «ОК». В ячейке B4 появился результат.

Захватив нижний правый угол ячейки B4, автозаполнением растягиваем результат на ячейки B4:B44 (см. рис. 1.6).

Аналогично находим значения функции $f(x)$ для *N*(2;1); *N*(0;0,05); *N*(1;2).

Строим график по данным. Выделяем диапазон ячеек A4:E44. Вызываем мастер диаграмм, выбрав пункты меню *Вставка*/*Диаграммы*. Выбираем тип диаграммы «*Точечная*» и вид – «*Точечная с гладкими кривыми*». Получаем графики плотности нормального распределения (см. рис. 1.6).

Делаем вывод: параметр *a* изменяет положение графика, с увеличением параметра график смещается вправо; параметр σ влияет на ширину графика, с увеличением параметра график растягивается.

Графики функции нормального распределения получим аналогично – при вызове функции НОРМ.РАСП в поле «Интегральная» нужно поставить 1.

	A	B	C	D	E	F	G	H	I	J
1	НОРМАЛЬНОЕ РАСПРЕДЕЛЕНИЕ									
2	Плотность распределения					Функция распределения				
3	x	N(0;1)	N(2;1)	N(0;0,5)	N(1;2)	x	N(0;1)	N(2;1)	N(0;0,5)	N(1;2)
4	-3	0,004	0,000	0,000	0,027	-3	0,001	0,000	0,000	0,023
5	-2,8	0,008	0,000	0,000	0,033	-2,8	0,003	0,000	0,000	0,029
6	-2,6	0,014	0,000	0,000	0,039	-2,6	0,005	0,000	0,000	0,036
7	-2,4	0,022	0,000	0,000	0,047	-2,4	0,008	0,000	0,000	0,045
8	-2,2	0,035	0,000	0,000	0,055	-2,2	0,014	0,000	0,000	0,055
9	-2	0,054	0,000	0,000	0,065	-2	0,023	0,000	0,000	0,067
10	-1,8	0,079	0,000	0,001	0,075	-1,8	0,036	0,000	0,000	0,081
11	-1,6	0,111	0,001	0,005	0,086	-1,6	0,055	0,000	0,001	0,097
12	-1,4	0,150	0,001	0,016	0,097	-1,4	0,081	0,000	0,003	0,115
13	-1,2	0,194	0,002	0,045	0,109	-1,2	0,115	0,001	0,008	0,136
14	-1	0,242	0,004	0,108	0,121	-1	0,159	0,001	0,023	0,159
15	-0,8	0,290	0,008	0,222	0,133	-0,8	0,212	0,003	0,055	0,184
16	-0,6	0,333	0,014	0,388	0,145	-0,6	0,274	0,005	0,115	0,212
17	-0,4	0,368	0,022	0,579	0,156	-0,4	0,345	0,008	0,212	0,242
18	-0,2	0,391	0,035	0,737	0,167	-0,2	0,421	0,014	0,345	0,274
19	0	0,399	0,054	0,798	0,176	0	0,500	0,023	0,500	0,309
20	0,2	0,391	0,079	0,737	0,184	0,2	0,579	0,036	0,655	0,345
21	0,4	0,368	0,111	0,579	0,191	0,4	0,655	0,055	0,788	0,382
22	0,6	0,333	0,150	0,388	0,196	0,6	0,726	0,081	0,885	0,421
23	0,8	0,290	0,194	0,222	0,198	0,8	0,788	0,115	0,945	0,460
24	1	0,242	0,242	0,108	0,199	1	0,841	0,159	0,977	0,500
25	1,2	0,194	0,290	0,045	0,198	1,2	0,885	0,212	0,992	0,540
26	1,4	0,150	0,333	0,016	0,196	1,4	0,919	0,274	0,997	0,579
27	1,6	0,111	0,368	0,005	0,191	1,6	0,945	0,345	0,999	0,618
28	1,8	0,079	0,391	0,001	0,184	1,8	0,964	0,421	1,000	0,655
29	2	0,054	0,399	0,000	0,176	2	0,977	0,500	1,000	0,691
30	2,2	0,035	0,391	0,000	0,167	2,2	0,986	0,579	1,000	0,726
31	2,4	0,022	0,368	0,000	0,156	2,4	0,992	0,655	1,000	0,758
32	2,6	0,014	0,333	0,000	0,145	2,6	0,995	0,726	1,000	0,788
33	2,8	0,008	0,290	0,000	0,133	2,8	0,997	0,788	1,000	0,816
34	3	0,004	0,242	0,000	0,121	3	0,999	0,841	1,000	0,841

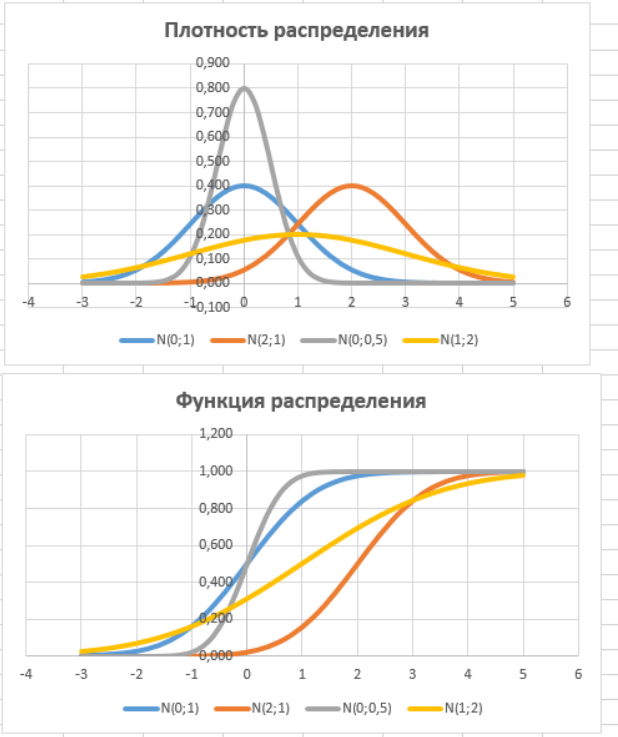


Рис. 1.6 – Часть исходных данных и решение контрольного примера 1.2

Контрольный пример 1.3. Исследование основных распределений в пакете Statistica.

1.3.1. *Основы работы в пакете Statistica*. Пакет Statistica состоит из статистических модулей для анализа и обработки данных, которые в свою очередь состоят из статистических процедур.

Выбор необходимого модуля осуществляется с помощью меню **Statistics**, либо с помощью кнопки в левом нижнем углу окна.

Прикладное окно пакета Statistica имеет стандартную для окна приложения MS Windows структуру (рис. 1.7).

Основным является рабочее окно, в котором вводятся исходные данные и выводятся результаты их статистической обработки в табличном или графическом виде. Ввод данных осуществляется в табличном виде

Набор данных в пакете Statistica – это прямоугольная таблица, столбцам которой соответствуют обрабатываемые *переменные* (*Variables*), а строкам отвечают *наблюдения* (**Cases**) значений переменных. Для создания нового набора данных нужно, прежде всего, завести файл с *трафаретом* таблицы нужных размеров, который открывается при запуске пакета (рис. 1.7).

Основным является рабочее окно, в котором вводятся исходные данные и выводятся результаты их статистической обработки в табличном или графическом виде. Ввод данных осуществляется в табличном виде.

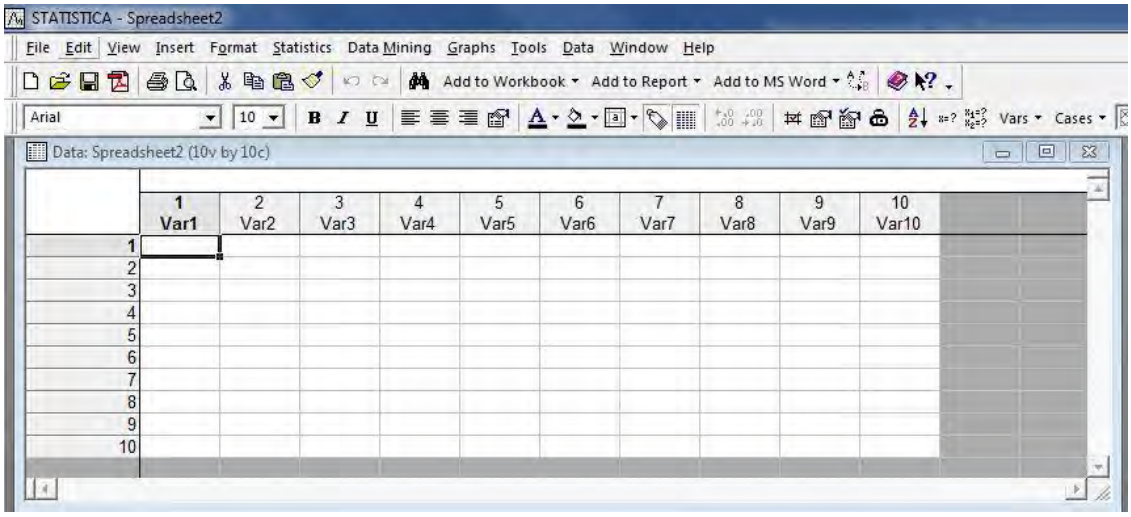


Рис. 1.7 – Окно пакета Statistica

Для работы с распределениями случайных величин в пакете Statistica используется калькулятор вероятностных распределений. Для его вызова выполняется процедура *Probability Calculator*:

Statistics – Probability Calculator – Distributions.

С помощью вероятностного калькулятора могут решаться разнообразные вероятностные задачи, например, построение графиков плотностей и функций распределения, определение квантили для заданной вероятности и пр.

В случае работы с дискретными распределениями для расчета вероятностей и функций распределения применяют встроенные функции.

В пакете Statistica имеются встроенные функции для всех основных законов распределения, причем функции пакета, имена которых начинаются с буквы *I*, вычисляют значения функций распределения.

Для вычисления встроенной функции в пакете Statistica следует выделить незаполненный столбец таблицы **Spreadsheet** исходных данных и затем выполнить команду *Variable Specs*, нажав правую кнопку мыши. В нижней части отрывшегося окна спецификации переменной находится рабочая *область Long name* (*label, link or formula with function*), которая предназначена для ввода выражений и комментариев. Набор формулы следует производить, начиная со знака равно (=), далее встроенные функции могут быть набраны

непосредственно с клавиатуры либо с помощью конструктора (кнопка *Function*).

1.3.2. *Нормальное распределение*.

Запустим процедуру *Probability Calculator*, для чего в меню *Statistics* выберем модуль *Probability Calculator*, выделим соответствующую строку и нажмем кнопку ОК. В результате откроется окно калькулятора *Probability Distribution Calculator*.

Выполним расчеты для нормального распределения со средним значением $a = 0{,}5$ и $\sigma = 1$. Первая задача будет заключаться в поиске квантили для вероятности $p = 0.8$ и построении графиков плотности и функции распределения.

В окне калькулятора в соответствующие поля введем параметры распределения, вероятность p и отметим опции *Inverse* и *Create Graph.*

После нажатия кнопки *Compute* получим результат (рис. 1.8). Значение квантили для заданной вероятности равно 1.341621.

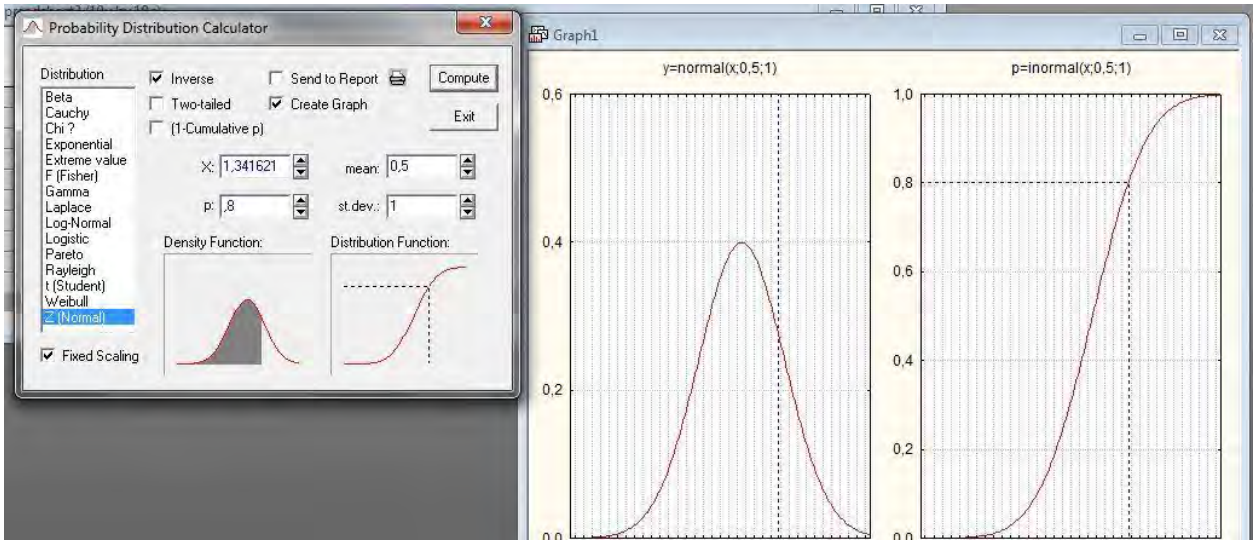


Рис. 1.8 – Графики плотности и функции распределения для нормального распределения

Следующая задача заключается в определении значения функции распределения для заданного значения случайной величины $x = 1$. Введя в поле X значение, равное 1, и нажав кнопку *Compute*, в поле p получим значение функции распределения, равное 0,691462 (рис. 1.9).

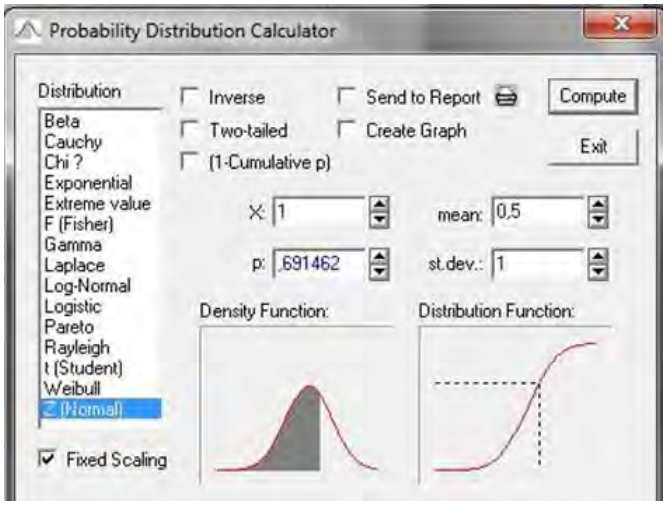


Рис. 1.9. Результаты расчета функции распределения для $x = 1$

1.3.3. Распределения χ^2, *Фишера и Стьюдента*.

Аналогично для распределения χ^2 с числом степеней свободы $df = 6$ и вероятностью $p = 0{,}8$ определена квантиль *Chi_I* = 8,558060 и построены графики плотности и функции распределения (рис. 1.10).

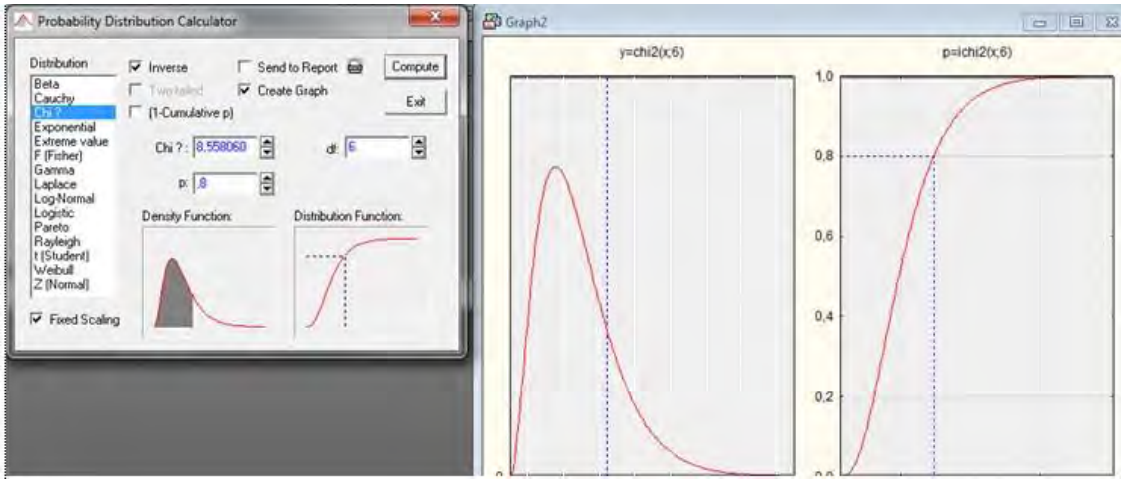


Рис. 1.10. Результаты расчета квантили и графики плотности
и функции распределения для распределения χ^2

Для распределения Фишера со степенями свободы $k_1 = 4$ и $k_2 = 15$ результаты расчетов представлены на рис. 1.11. В поля $df1$ и $df2$ введены соответственно значения k_1 и k_2.

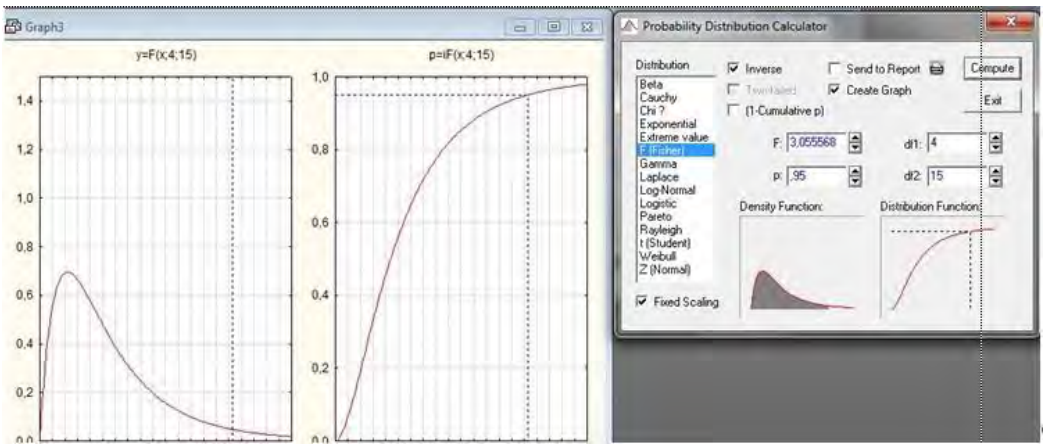


Рис. 1.11. Результаты расчета квантили и графики плотности и функции распределения для распределения Фишера

Для распределения Стьюдента с числом степеней свободы $df = 10$ результаты расчетов представлены на рис. 1.12. Квантиль этого распределения, соответствующая вероятности $p = 0{,}95$, равна $t = 1{,}812461$.

Графики плотности вероятности и функции распределения представлены на рис. 1.12.

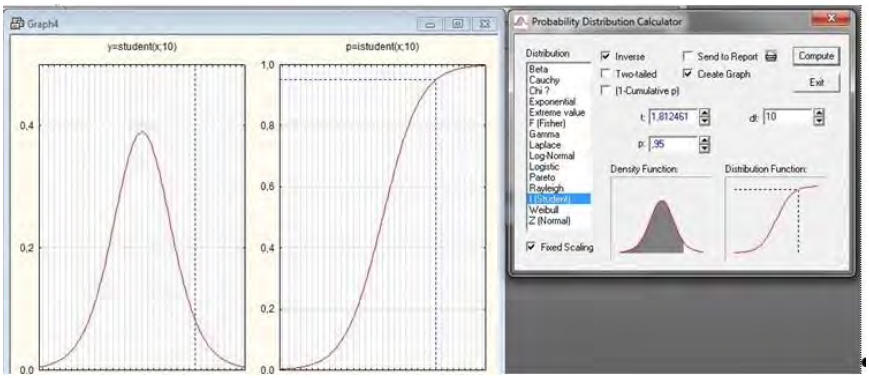


Рис. 1.12. Результаты расчета квантили и графики плотности и функции распределения для распределения Стьюдента

Задания для самостоятельной работы

Задание 1**.** Дан ряд распределения дискретной случайной величины X.

1) Используя Excel, найти числовые характеристики случайной величины X: математическое ожидание $M(X)$, дисперсию $D(X)$, среднее квадратическое отклонение $\sigma(X)$, коэффициент асимметрии As и эксцесс Ek; указать их размерность; построить многоугольник распределения случайной величины.

2) Используя найденные числовые характеристики, описать соответствующие им особенности распределения случайной величины: центр распределения, степень рассеивания значений случайной величины около центра распределения, степень скошенности, островершинность многоугольника распределения по сравнению с нормальной кривой

Число дорожно-транспортных происшествий в регионе за сутки является случайной величиной X, распределенной по закону:

1	X	0	1	2	3	4	5	6	7	8	9
	p	0,011	0,081	0,106	0,111	0,295	0,161	0,097	0,075	0,036	0,027
2	X	0	1	2	3	4	5	6	7	8	9
	p	0,023	0,031	0,057	0,075	0,12	0,323	0,136	0,093	0,087	0,054

Успеваемость студентов (баллы) по результатам аттестации является случайной величиной X, распределенной по закону:

3	X	0	0,5	1	1,5	2	2,5	3	3,5	4	4,5	5
	p	0,011	0,014	0,02	0,097	0,101	0,135	0,139	0,235	0,106	0,089	0,053
4	X	0	0,5	1	1,5	2	2,5	3	3,5	4	4,5	5
	p	0,01	0,013	0,023	0,097	0,105	0,129	0,138	0,236	0,016	0,087	0,056

Количество осадков (мм), выпавших за сезон, является случайной величиной X, распределенной по закону:

5	X	0	2,5	3	3,5	4	4,5	5	5,5	6	6,5	7
	p	0,008	0,06	0,096	0,177	0,210	0,135	0,109	0,085	0,056	0,039	0,025
6	X	0	2,5	3	3,5	4	4,5	5	5,5	6	6,5	7
	p	0,007	0,05	0,096	0,167	0,22	0,135	0,124	0,086	0,054	0,041	0,02

Количество деталей, обработанных на станке за 1 час, является случайной величиной X, распределенной по закону:

5	X	0	1	2	3	4	5	6	7	8	9	10
	p	0,009	0,033	0,051	0,069	0,073	0,084	0,148	0,223	0,187	0,09	0,035
6	X	0	1	2	3	4	5	6	7	8	9	10
	p	0,008	0,033	0,051	0,067	0,073	0,084	0,148	0,223	0,187	0,09	0,036

Число соединений в минуту на автоматической телефонной станции является случайной величиной X, распределенной по закону:

7	X	0	1	2	3	4	5	6	7	8	9	10
	p	0,012	0,025	0,029	0,069	0,175	0,2	0,168	0,136	0,09	0,087	0,009
8	X	0	1	2	3	4	5	6	7	8	9	10
	p	0,01	0,023	0,033	0,067	0,173	0,214	0,169	0,116	0,097	0,087	0,011

Число машин, проезжающих через данный перекресток за 1 минуту, является случайной величиной X, распределенной по закону:

9	X	0	1	2	3	4	5	6	7	8	9	10
	p	0,008	0,043	0,051	0,067	0,085	0,108	0,113	0,185	0,198	0,066	0,076
10	X	0	1	2	3	4	5	6	7	8	9	10
	p	0,009	0,043	0,051	0,069	0,083	0,104	0,12	0,183	0,216	0,087	0,035

Содержание минерала A (%) в образцах является случайной величиной X, распределенной по закону:

11	X	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1	1,1
	p	0,039	0,063	0,085	0,109	0,306	0,184	0,094	0,061	0,035	0,016	0,008
12	X	0,1	0,2	0,3	0,4	0,5	0,6	0,7	0,8	0,9	1	1,1
	p	0,039	0,06	0,084	0,111	0,304	0,185	0,091	0,065	0,033	0,017	0,011

Задание 2. В пакете Excel*:*

1. Ввести в таблицу значения аргумента x в диапазоне от $x1$ до $x2$ с шагом h.

2. Вычислить значение плотности стандартного нормального распределения, а также плотности нормального распределения с параметрами a и σ (табл. 1.2).

3. Для заданных параметров нормального распределения построить семейство графиков функции распределения и плотности распределения вероятностей.

4. В пакете Statistica, используя вероятностный калькулятор, рассчитать квантили нормального распределения с параметрами a и σ из табл. 1.2 для $p =$ 0,8; 0,9; 0,95; 0,99, а также построить графики $f(x)$ и $F(x)$

Таблица 1.2

Вариант	$[x_1, x_2]$	h	a_1	σ_1	a_2	σ_2	a_3	σ_3
1	[–1; 7]	0,5	2	3	0	0,8	3	1
2	[0; 2]	0,1	1	1	0	2	2	3,2
3	[2; 4]	0,1	1	3	2	0,1	3	0,6
4	[–4; 4]	0,4	0	2	1	1,5	4	2
5	[0,3; 8,3]	0,4	4	0,2	1	0,1	3	2
6	[–2; 4]	0,3	1	0,7	0	2	2	0,5
7	[–4; 7,4]	0,6	1	2	0,5	0,5	2	2,8
8	[1; 3]	0,1	2	0,5	1	2	0	22
9	[–1; 3]	0,2	0,2	1	2	0,5	0	2,8
10	[0,3; 8,3]	0,4	4	0,2	1	3,5	1	4
11	[0; 10]	0,5	5	2	3	0,8	2	5
12	[0; 4]	0,2	2	0,8	1,5	1	1	2,5

Задание 3. В пакете Statiistica по данным, приведенным в таблице 1.3:

1) Построить график плотности распределения а) хи-квадрат, б) Стьюдента, протабулировав эту функцию на отрезке от x_1 до x_2 с шагом h и взяв степень свободы k_1. Проанализировать зависимость параметра распределения k_1 на график.

2) Построить график плотности распределения Фишера, протабулировав эту функцию x_1 до x_2 с шагом и взяв степени свободы k_1 и k_2. Проанализировать зависимость параметров распределения k_1 и k_2 на график.

Таблица 1.3

Вариант	x_1	x_2	h	k_1	k_2	**Вариант**	x_1	x_2	h	k_1	k_2
1	0	5	0,2	4	5	**7**	0	4	0,15	5	7
2	0	6	0,2	4	6	**8**	0	6	0,2	5	6
3	0	4	0,1	5	6	**9**	0	5	0,3	4	6
4	0	7	0,2	6	7	**10**	0	10	0,3	6	7
5	0	5	0,1	4	6	**11**	0	3	0,05	5	6
6	0	8	0,2	3	5	**12**	0	8	0,3	5	6

Лабораторная работа № 2. Первичная обработка статистических данных. Точечные и интервальные оценки характеристик случайной величины

Необходимые теоретические сведения для выполнения лабораторной работы находятся в теоретическом разделе – тема 2.1.

Контрольный пример 2.1. Дана выборка значений массы тела учащихся в килограммах:

Эмпирические данные о массе тела учащихся

Наблюдения				
64	62	58	58	61
57	62	63	63	58
63	60	61	61	60
62	64	59	59	64
58	61	62	62	60
61	59	60	60	59
63	59	60	60	61
60	63	58	58	64
60	61	61	61	62
61	62	60	60	59
65	58	63	63	65

Необходимо в пакетах *Statistica* и *Excel*:

1) Построить дискретный вариационный ряд, провести интервальную обработку. Построить полигон частот, гистограмму относительных частот и кумулятивную кривую.

2) рассчитать выборочные характеристики распределения. Найти 95 % доверительный интервал для математического ожидания и дисперсии. Определить, сколько нужно иметь наблюдений, чтобы точность определения математичекого ожидания не превышала 0,2.

Решение.

1. Упорядочим выборку, т. е. составим вариационный ряд:

57	57	58	58	58	58	58	58	59	59	59	59
59	59	59	60	60	60	60	60	60	60	60	60
60	61	61	61	61	61	61	61	61	61	62	62
62	62	62	62	62	62	63	63	63	63	63	63
63	64	64	64	64	65	65					

Дискретный вариационный ряд имеет вид табл. 2.1.

Таблица 2.1

x_i	57	58	59	60	61	62	63	64	65	$\sum$
n_i	2	6	7	10	9	8	7	4	2	**55**
w_i	$\frac{2}{55}$	$\frac{6}{55}$	$\frac{7}{55}$	$\frac{10}{55}$	$\frac{9}{55}$	$\frac{8}{55}$	$\frac{7}{55}$	$\frac{4}{55}$	$\frac{2}{55}$	**1**

Проведем интервальную обработку.

По условию, объем выборки $n = 55$. Определим оптимальную длину частичного интервала:

$$h = \frac{x_{\max} - x_{\min}}{l} = \frac{x_{\max} - x_{\min}}{1 + 3{,}322 \lg n} = \frac{65 - 67}{1 + 3{,}322 \lg 55} \approx \frac{8}{7} \approx 1{,}15.$$

Тогда:

$x_1 = 57; x_2 = 57 + 1{,}15 = 58{,}15; x_3 = 58{,}15 + 1{,}15 = 59{,}3;$

$x_4 = 59{,}3 + 1{,}15 = 60{,}45; x_5 = 60{,}45 + 1{,}15 = 59{,}3;$

$x_6 = 61{,}6 + 1{,}15 = 62{,}75 x_7 = 62{,}75 + 1{,}15 = 63{,}9;\ x_8 = 63{,}9 + 1{,}15 = 65{,}05.$

Тогда интервальный статистический ряд примет вид табл. 2.2.

Таблица 2.2

x_i	x_{i+1}	x_i^*	n_i	$\frac{n_i}{n}$	Накопленная частота
57	58,15	58,575	8	8/55	8/55
58,15	59,3	58,725	7	7/55	15/55
59,3	60,45	59,875	10	10/55	25/55
60,45	61,6	61,025	9	9/55	34/55
61,6	62,75	62,175	8	8/55	42/55
62,75	63,9	63,325	7	7/55	49/55
63,9	65,05	64,475	6	6/55	1
$\sum$			**55**	**1**	

2. В пакете *Excel* в ячейку A1 введем слово «Наблюдения» (рис. 2.1), а в диапазон A2:A56 – исходные данные.

	A	B	C	D	E	F	G	H
1	Наблюдения		Максимум	65	Варианты	Абсолютные частоты	Относительные частоты	Накопленные частоты
2	64		Минимум	57	57	2	0,0364	0,0364
3	57				58	6	0,1091	0,1455
4	63				59	7	0,1273	0,2727
5	62				60	10	0,1818	0,4545
6	58				61	9	0,1636	0,6182
7	61				62	8	0,1455	0,7636
8	63				63	7	0,1273	0,8909
9	60				64	4	0,0727	0,9636
10	60				65	2	0,0364	1,0000
11	61				Всего наблюдений	55	1	
12	65							
13	62							

Рис. 2.1. Результат вычислений абсолютных, относительных и накопленных частот

Рассчитаем максимальное и минимальное значения выборочных данных в ячейках D1 и D2, введя соответственно функции МАКС(A2:A56) и МИН(A2:A56) (рис. 2.1).

Построим вариационный ряд, считая массу тела дискретной случайной величиной. В ячейку E1 введем заголовок «Варианты», а ниже в столбце – все возможные неповторяющиеся значения массы тела учащихся (x_i), которые встречались в выборке (от минимального до максимального).

В ячейке F1 запишем заголовок «Абсолютные частоты». В этом столбце будут рассчитаны значения частот n_i, т. е. то количество раз, которое соответствующее значение x_i случайной величины встречалось в выборке. Для заполнения столбца абсолютных частот можно использовать стандартную функцию ЧАСТОТА().

Выделим мышью диапазон F2:F10, в котором разместятся найденные частоты, вызовем *Мастер функций* и в категории *Статистические* выберем функцию ЧАСТОТА. После этого заполним ее аргументы:

– *массив данных* – это диапазон эмпирических данных A2:A56;

– *массив интервалов* – это диапазон значений вариант E2:E10.

Закончить ввод функции нужно одновременным нажатием клавиш *Ctrl* + *Shift* + *Enter*, поскольку ее результатом является диапазон значений. В строке формул эта функция будет показана в фигурных скобках.

В ячейке F11 найдем общее число наблюдений, просуммировав значения в столбце абсолютных частот (см. рис. 2.1).

В ячейке G1 запишем заголовок «*Относительные частоты*». Для расчета относительных частот $w_i = \frac{n_i}{n}$ внесем в ячейку G2 формулу $=$F2/$F$11

и скопируем ее методом автозаполнения вниз по столбцу. Сумма относительных частот в этом столбце должна быть равна единице.

Последний столбец таблицы озаглавим «*Накопленные частоты*». В ячейку H2 скопируем значение относительной частоты из ячейки G2, а в ячейку H3 введем формулу $=H2+G3$. Методом автозаполнения скопируем введенную формулу вниз по столбцу в диапазон H4:H10.

Итоговый вид таблицы после форматирования показан на рис. 2.1.

Построим полигон частот по данным в столбце «Абсолютные частоты», как показано на рис. 2.2 (используем диаграмму типа «точечная с прямыми отрезками и маркерами»).

Рис. 2.2. Полигон частот

Построим также совместную диаграмму относительных и накопленных частот. Чтобы совместить в диаграмме несколько типов, (например, Гистограмму и Линейный график – как в нашем примере), необходимо сначала построить все диаграммы одного вида. Выделим диапазон G2:H10 и на вкладке «*Вставка*» выберем тип *Гистограмма* (рис. 2.3 – а).

Теперь выбираем один ряд и для него меняем тип диаграммы. Щелкнув на ряде 2 правой кнопкой мыши, выбираем «*Изменить тип диаграммы для ряда*» и тип «*График*» (рис. 2.3 – б).

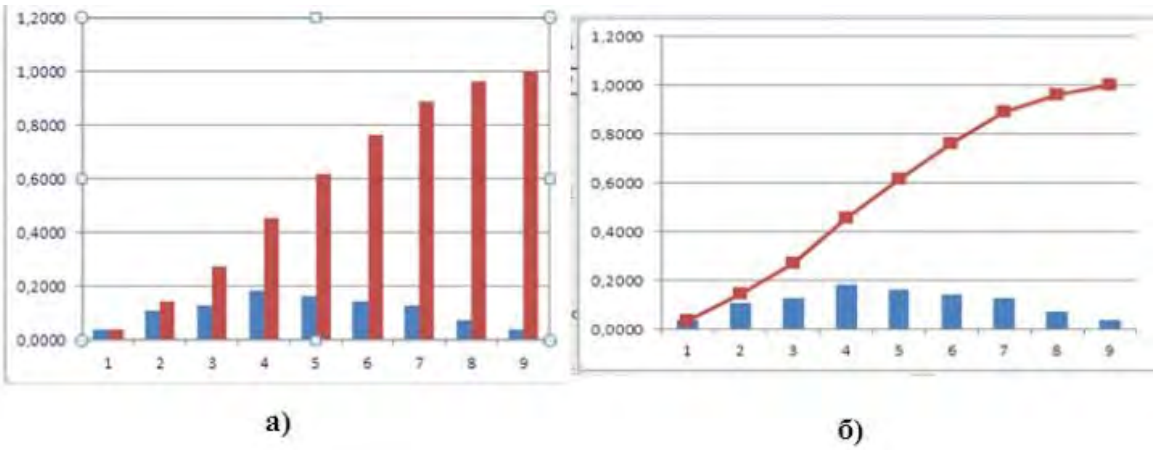


Рис. 2.3. – Первый этап построения диаграммы относительных и накопленных частот

Так как гистограмма получилась незаметной на диаграмме, нужно добавить вспомогательную ось. Для этого нажмем правой кнопкой мыши на гистограмму или на название в легенде. Далее в появившимся диалоговом окне выберем «*Формат ряда данных»*. В открывшемся окне ищем *Параметры ряда* и меняем галочку на «*По вспомогательной оси»*. После минимального редактирования диаграмма будет иметь такой вид, как показано на рис. 2.4.

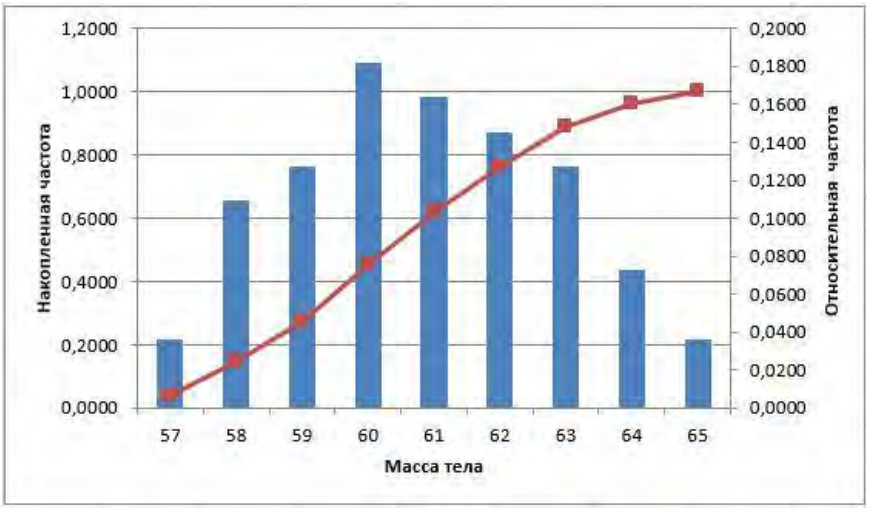


Рис. 2.4. Диаграмма относительных и накопленных частот

Гистограмма относительных частот есть аналог функции плотности распределения, а график накопленных частот проходит через левые концы «ступенек» эмпирической функции распределения и называется *кумулятивной кривой*.

Для построения интервального вариационного ряда разобьем диапазон наблюдавшихся значений [57; 65] на интервалы шириной 1,15. При этом минимальное значение должно попасть внутрь интервала. Данные в столбце «*Варианты*» (рис. 2.5) интерпретируются как правые границы интервалов. Значения, которые дает функция ЧАСТОТА – это частоты попадания в интервал. При этом если значение случайной величины попадает на границу интервала, то оно учитывается в левом интервале. Что касается самого первого значения в столбце «*Варианты*», то для него функция ЧАСТОТА дает количество наблюдений меньших или равных ему.

Остальные расчеты полностью аналогичны. На рис. 2.5 показан результат вычисления абсолютных, относительных и накопленных частот для интервального ряда.

F2 {=ЧАСТОТА(A2:A56;E2:E8)}

	A	B	C	D	E	F	G	H
1	Наблюдения		Максимум	65	Варианты	Абсолютные частоты	Относительные частоты	Накопленные частоты
2	64		Минимум	57	58,15	8	0,1455	0,1455
3	57				59,3	7	0,1273	0,2727
4	63				60,45	10	0,1818	0,4545
5	62				61,6	9	0,1636	0,6182
6	58				62,75	8	0,1455	0,7636
7	61				63,9	7	0,1273	0,8909
8	63				65,05	6	0,1091	1,0000
9	60			Всего наблюдений		55	1	

Рис. 2.5. Результаты расчетов частот для интервального ряда

В пакете *Microsoft Excel* для определения выборочных оценок параметров распределения используются следующие функции:

СРЗНАЧ – вычисляет среднюю арифметическую аргументов (т. е. выборочную среднюю);

МЕДИАНА – находит медиану заданной выборки;

МОДА.ОДН – вычисляет наиболее часто встречающееся в выборке значение;

ДИСП.Г – вычисляет выборочную дисперсию;

ДИСП.В – вычисляет «исправленную» дисперсию;

СТАНДАРТОТКЛОН.В – вычисляет «исправленное» СКО;

ЭКСЦЕСС – вычисляет оценку эксцесса по выборке;

СКОС – позволяет оценить асимметрию выборочного распределения.

Кроме того, в надстройке *Пакет анализа* имеется инструмент *Описательная статистика*, который дает возможность получить все выборочные характеристики случайной величины.

Введем эмпирические данные о весе учащихся на чистый лист *Excel* и оформим его, как показано на рис. 2.6.

В ячейки D2:D9 для определения выборочных числовых характеристик введем стандартные функции Excel категории *Статистические*. Аргументами всех этих функций является диапазон выбранных значений веса A2:A56.

	A	B	C	D	E	F	G	H
1	Наблюдения		Выборочные оценки (используя стандартные функции)				Выборочные оценки (пакет анализа)	
2	64		Среднее	60,855			*Наблюдения*	
3	57		Выборочная дисперсия	4,124				
4	63		Исправленная дисперсия	4,201			Среднее	60,855
5	62		Стандартное отклонение	2,050			Стандартная ошибка	0,276
6	58		Мода	60			Медиана	61
7	61		Медиана	61			Мода	60
8	63		Эксцесс	-0,744			Стандартное отклонение	2,050
9	60		Асимметрия	0,096			Дисперсия выборки	4,201
10	60		Количество наблюдений	55			Эксцесс	-0,744
11	61						Асимметричность	0,096
12	65						Интервал	8
13	62						Минимум	57
14	62						Максимум	65
15	60						Сумма	3347
16	64						Счет	55
17	61							

Рис. 2.6. Лист *Excel* с расчетом выборочных характеристик распределения

Распределение веса студентов является достаточно симметричным (асимметрия равна 0,096 и близка к нулю), а эксцесс имеет небольшое отрицательное значение (–0,744). Это означает, что распределение веса имеет более низкую и пологую вершину по сравнению с нормальным распределением.

Аналогичные данные можно получить с помощью инструмента *Описательная статистика* надстройки *Пакет Анализа*. Зададим команду *Данные/Анализ данных* и выберем инструмент *Описательная статистика*. Заполним диалоговое окно, как показано на рис. 2.7.

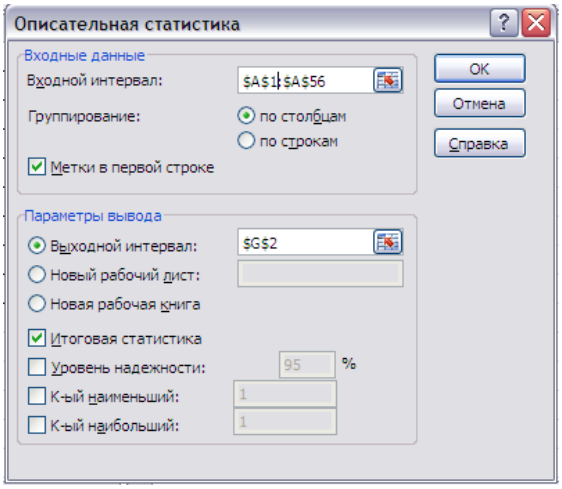


Рис. 2.7. Диалоговое окно для вывода описательной статистики

Данные каждой выборки должны быть расположены в одном столбце или одной строке. Переключатель *Группирование* в нашем случае установлен в положение по столбцам, так как эмпирические данные занесены в первый столбец.

Флажок *Метки в первой строке* установлен, поскольку входной интервал включает и заголовок данных в первой строке (слово «Наблюдения»).

Переключатель *Параметры вывода* установлен в положение *Выходной интервал*, так как нужно получить результаты расчетов на текущем листе *Excel*. В соответствующем поле указан адрес левой верхней ячейки выходного диапазона (G2).

Установим значок *Итоговая статистика* для вывода на листе Excel выборочных характеристик.

После заполнения этого диалогового окна нажимаем кнопку *OK*. Результаты расчетов с помощью *Пакета анализа* показаны на рис. 2.6 в столбцах G и H.

В пакете *Excel* существуют два основных варианта нахождения доверительного интервала для математического ожидания: когда дисперсия известна и когда неизвестна. В первом случае для вычислений применяется функция ДОВЕРИТ.НОРМ, а во втором – ДОВЕРИТ.СТЮДЕНТ.

Скопируем исходные данные на чистый лист *Excel* (см. рис. 3.3) и рассчитаем с помощью стандартных функций основные числовые характеристики: выборочное среднее, выборочное и «исправленное» стандартное отклонение.

Рассчитаем доверительный интервал с помощью функции ДОВЕРИТ.НОРМ. Ее синтаксис:

= ДОВЕРИТ.НОРМ (альфа; стандартное_откл; размер).

Здесь «альфа» – уровень значимости; «стандартное_откл» – стандартное отклонение предлагаемой выборки; «размер» – объем выборки.

Граница доверительного интервала определяется по формуле:

$\bar{x} \pm$ ДОВЕРИТ.НОРМ.

Выделим ячейку, куда будет выводиться результат обработки данных (рис. 3.3 – ячейка H3). Щелкаем по кнопке «*Вставить функцию*».

Появляется *Мастер функций*. Переходим в категорию «*Статистические*» и выделяем «ДОВЕРИТ.НОРМ». Нажимаем *ОК*.

Открывается окно аргументов. Заполняем поля следующим образом: =ДОВЕРИТ.НОРМ(0,05;D2;55).

В ячейке D2 находится значение σ, рассчитанное с помощью стандартной функции Excel: =СТАНДОТКЛОН.Г(A1:A55).

Производим расчет левой границы доверительного интервала. Для этого выделяем отдельную ячейку (H4), ставим знак «=» и вычитаем содержимое элементов листа, в которых расположены результаты вычислений выборочного среднего (ячейка D1) и ДОВЕРИТ.НОРМ. В нашем случае получилась следующая формула: =D1 – H3.

Таким же образом производим вычисление правой границы доверительного интервала: =D1 + H3 (см. рис. 2.8).

Доверительный интервал для математического ожидания, найденный с помощью фунции ДОВЕРИТ.НОРМ, имеет вид: $60{,}319 < a < 61{,}391$.

Функция ДОВЕРИТ.СТЮДЕНТ выполняет вычисление доверительного интервала генеральной совокупности с использованием распределения Стьюдента. Его удобно использовать в том случае, когда дисперсия и, соответственно, стандартное отклонение неизвестны. Синтаксис оператора таков:

=ДОВЕРИТ.СТЮДЕНТ(альфа; стандартное_откл; размер).

Рассчитаем границы доверительного интервала с неизвестным стандартным отклонением на примере нашей выборочной совокупности.

Выделим ячейку, в которую будет производиться расчет (H7 – рис. 2.8). Щелкаем по кнопке *«Вставить функцию»*.

В открывшемся *Мастере функций* переходим в категорию *«Статистические»*. Выбираем наименование «ДОВЕРИТ.СТЬЮДЕНТ» и нажимаем *ОК*.

Заполняем поля: =ДОВЕРИТ.СТЬЮДЕНТ(0,05;D3;55).

В ячейке D3 находится «исправленное» стандартное отклонение, вычисленное с помощью функции *Excel* СТАНДОТКЛОН.В(A1:A55).

Далее, в ячейках H8 и H9 рассчитываем левую и правую границы доверительного интервала по формуле D1 μ H7 (см. рис. 2.8).

	A	C	D	F	G	H	I	J	K
1	64	Среднее	60,855	Расчет доверительного интервала для M(X)					
2	57	Выборочное СКО	2,031	ДОВЕРИТ.НОРМ					
3	63	Исправленное СКО	2,050	Точность		0,537			
4	62	Доверительноя вероятность	0,95	Левая граница		60,318			
5	58			Правая граница		61,391			
6	61	Расчет доверительного интервала для D(X)		ДОВЕРИТ.СТЮДЕНТ					
7	63	Верхний Хи2-квантиль	76,192	Точность		0,554			
8	60	Нижний Хи2-квантиль	35,586	Левая граница		60,300			
9	60	Левая граница интервала	2,977	Правая граница		61,409			
10	61	Правая граница интервала	6,374						
11	65			Количество наблюдений n для точности 0,2					
12	62					212			

Рис. 2.8. Расчет доверительных интервалов для параметров случайной величины

Доверительный интервал, найденный с помощью функции ДОВЕРИТ.СТЬЮДЕНТ: $60{,}3 < a < 61{,}409$.

Полученная точность доверительного интервала $\delta \approx 0{,}55$ превышает заданное значение 0,2. Определим по формуле (2.9), сколько необходимо иметь данных наблюдений для достижения этой точности. Для этого внесем, например, в ячейку H12 следующую формулу:

=СТЬЮДЕНТ.ОБР.2Х(0,05;54)*(D3^2)/(0,2^2)+1.

Полученное значение показывает, что нужно иметь не менее 212 наблюдений.

Для нахождения доверительного интервала для дисперсии рассчитаем квантили хи-квадрат распределения с помощью стандартной функции пакета ХИ2.ОБР.ПХ:

Верхний χ^2 квантиль: =ХИ2.ОБР.ПХ(0,05/2;54).

Нижний χ^2 квантиль: =ХИ2.ОБР.ПХ(1-0,05/2;54).

Границы доверительного интервала для дисперсии, рассчитанные по формуле 2.8, на рис. 2.8 находятся в ячейках D9 и D10.

Запишем доверительный интервал для дисперсии: $2{,}997 < \sigma^2 < 6{,}374$.

Доверительный интервал для стандартного отклонения может быть получен путем извлечения квадратного корня из вышеуказанного выражения: $1{,}7254 < \sigma < 2{,}5247$.

3. Теперь выполним данные задания в пакете *Statistica*.

После запуска пакета на экране появится сетка-таблица.

Преобразуем таблицу к размерам 1×55, выполнив следующие действия.

Нажимаем кнопку *Vars* (на экране), в раскрывающемся меню выбираем *Delete*; появится окно *Delete Variables*.

Укажем, какие переменные-столбцы убрать.

Нажимаем кнопку *Cases*, выбираем опцию *Add* (добавление), появится окно *Add Cases*: укажем, сколько строк добавить и куда.

Далее выделим столбец – переменную *Var*1 (щелчком правой кнопки мыши по ее заглавию) – в открывшемся меню выберем *Variable Specs* (спецификации переменной) – в появившемся окне *Variable 1* введем *Name X*. Зададим исходные данные (или скопируем из пакета *Excel*).

В меню *Statistics – Basic Statistics/Tables* в окне *Descriptive Statistic* во вкладке *Quick* нажмем на кнопку *Frequency Table*. В результате получим таблицу частот (рис. 2.9). В первом столбце заданы интервалы для переменной *X*, причем последняя строка содержит пропущенные значения. Второй столбец содержит число попаданий переменной в интервалы (*Count*), третий столбец – накопленную частоту (*Cumulative Count*), четвертый и шестой – частоты в процентных соотношениях для имеющихся в наличии (не пропущенных) наблюдений (*Percent of Valid*) и для всех наблюдений (*% of Cases*), пятый и седьмой столбцы – накопленные частоты в процентах соответственно для (не пропущенных) наблюдений (*Cumul. % of Valid*) и для всех наблюдений (*Cumul. % of All*).

Frequency table: X (Spreadsheet1)
K-S d=,11619, p> ,20; Lilliefors p<,10

Category	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All
56,00000<x<=57,00000	2	2	3,63636	3,6364	3,63636	3,6364
57,00000<x<=58,00000	6	8	10,90909	14,5455	10,90909	14,5455
58,00000<x<=59,00000	7	15	12,72727	27,2727	12,72727	27,2727
59,00000<x<=60,00000	10	25	18,18182	45,4545	18,18182	45,4545
60,00000<x<=61,00000	9	34	16,36364	61,8182	16,36364	61,8182
61,00000<x<=62,00000	8	42	14,54545	76,3636	14,54545	76,3636
62,00000<x<=63,00000	7	49	12,72727	89,0909	12,72727	89,0909
63,00000<x<=64,00000	4	53	7,27273	96,3636	7,27273	96,3636
64,00000<x<=65,00000	2	55	3,63636	100,0000	3,63636	100,0000
Missing	0	55	0,00000		0,00000	100,0000

Рис. 2.9 – Таблица частот

Для построения графика частот (полигона частот) выделим столбец *Percent of valid*, нажмем правую кнопку мыши и в контекстном меню выберем команду *Graph of Block Data – Line Plot: Entire Columns* (так как данные для построения графика расположены в столбцах). В результате получим график, представленный на рис. 2.10.

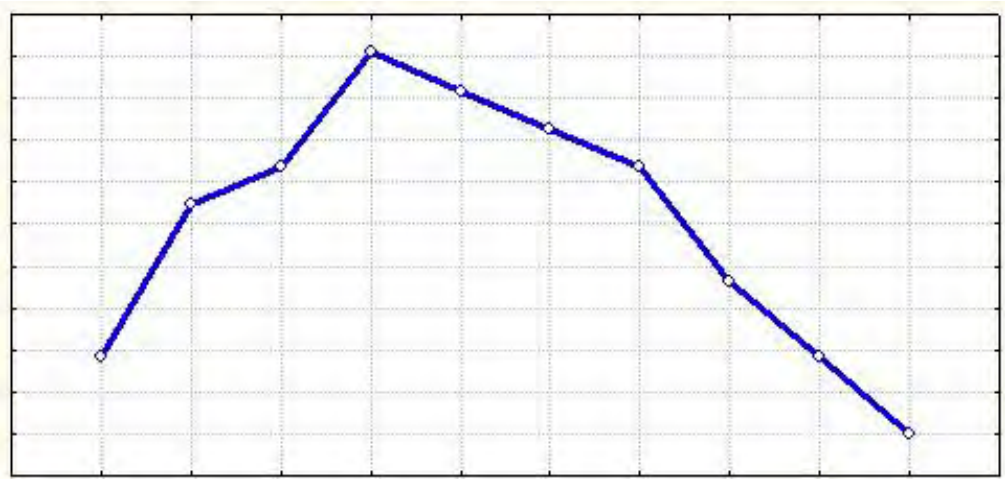


Рис. 2.10 – График, построенный по исходным данным

Для построения гистограммы частот во вкладке *Quick* окна *Descriptive Statistic* нажмем на кнопку *Histograms*. Получим гистограмму, представленную на рис. 2.11.

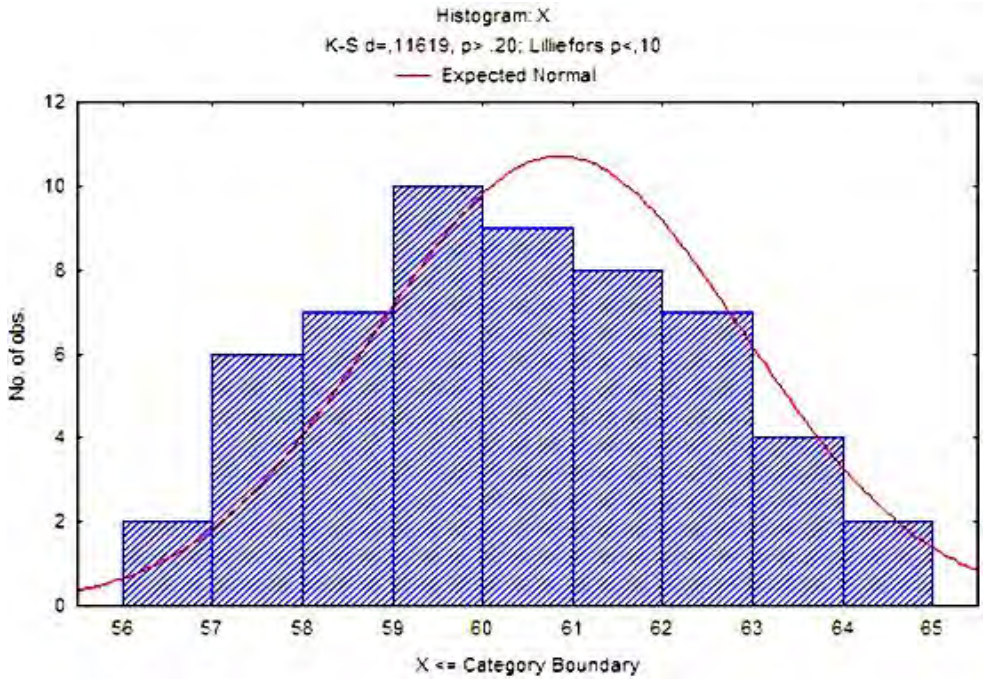


Рис. 2.11 – Гистограмма, построенная по исходным данным

Для построения кумулятивной кривой (эмпирической функции распределения) перейдем в окно *Graphs* на вкладку *Histograms*. Во вкладке *Advanced* выберем *Fit type – Off*, установим: *Graph type: Regular, Showing type: Cumulative*, *Variables – X*; *Categories (число интервалов группировки) – 250 – OK*.

Наблюдаем график кумулянты (рис. 2.12).

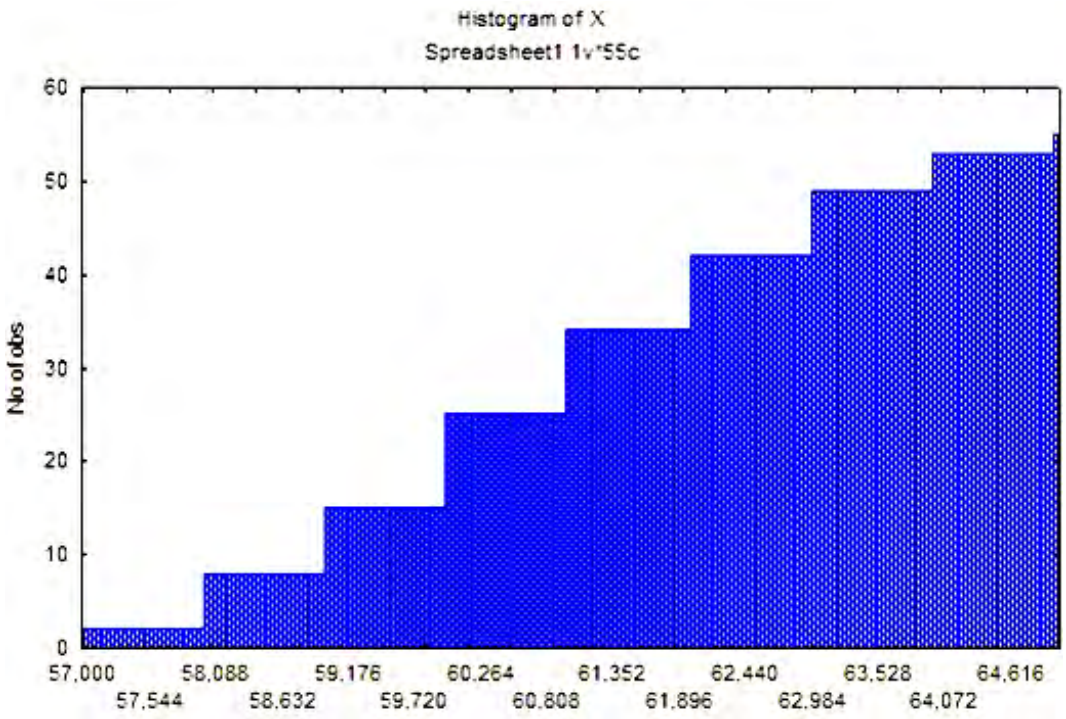


Рис. 2.12. Кумулятивная кривая

Для нахождения точечных и интервальных оценок будем использовать исходную выборку (столбец) из 55 наблюдений над случайной величиной. Скопируем данные с листа *Excel* в *Statistica*, предварительно создав в пакете новый документ.

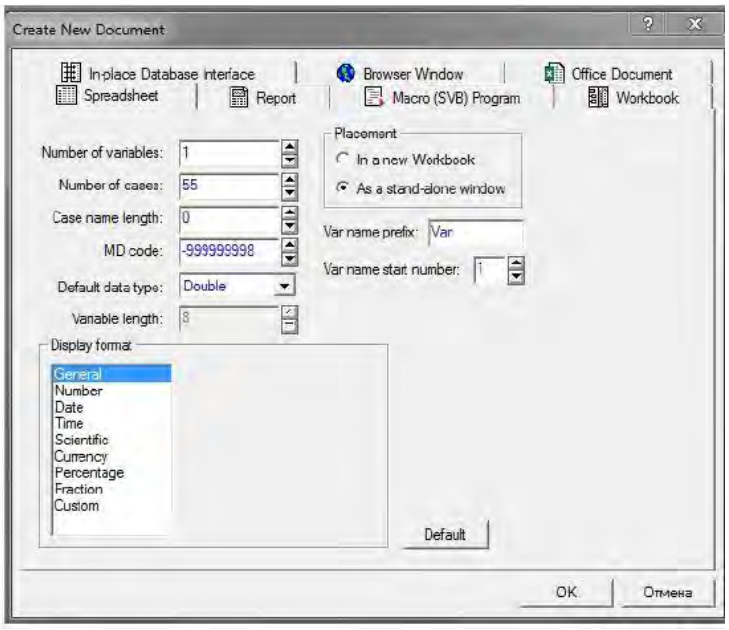


Рис. 2.13. Создание нового документа в пакете *Statistica*

В меню *Statistics – Basic Statistic/Tables* в окне *Descriptive Statistic* выберем вкладку *Advanced*, в результате появится окно, содержащее список числовых характеристик, которые могут быть вычислены. Отметим: *Mean* (среднее), *Median* (медиана), *Mode* (мода), *Standard Deviation* (среднее квадратиче-

ское отклонение), *Variance* (дисперсия), *Std.err. of mean (*стандартная ошибка среднего*), Skewness* (асимметрия), *Kurtosis* (эксцесс), *Range* (размах варьирования) – и активизируем кнопку *Summary*. В появившемся окне выделим переменную, для которой нужно произвести расчеты. В данном случае это переменная *X*.

Результаты вычислений размещаются в активном окне внизу (при анализе данных столбца) или слева (при анализе данных строки) от исходных данных (рис. 2.14).

Descriptive Statistics (Spreadsheet1)

Variable	Mean	Median	Mode	Frequency of Mode	Range	Variance	Std Dev.	Standard Error	Skewness	Kurtosis
X	60,85455	61,00000	60,00000	10	8,000000	4,200673	2,049554	0,276362	0,096321	-0,743698

Рис. 2.14 – Вычисленные параметры описательной статистики

Предположение о характере генерального распределения. Так как для нормального распределения эксцесс и асимметрия равны нулю, то достаточно малые значения эксцесса и асимметрии, а также то, что мода и медиана близки к среднему выборочному, позволяют выдвинуть гипотезу о нормальном распределении генеральной совокупности.

Определим доверительные интервалы для a и σ с уровнем доверия $\gamma = 0{,}95$.

Возвращаемся на вкладку *Advanced* процедуры *Descriptive Statistics*. Установим *Conf. Limits for means, Cl for Sample SD* и укажем значение *Interval: 95 %* (см. рис. 2.15), нажав на кнопку *Variables*, зададим имя переменной *X*.

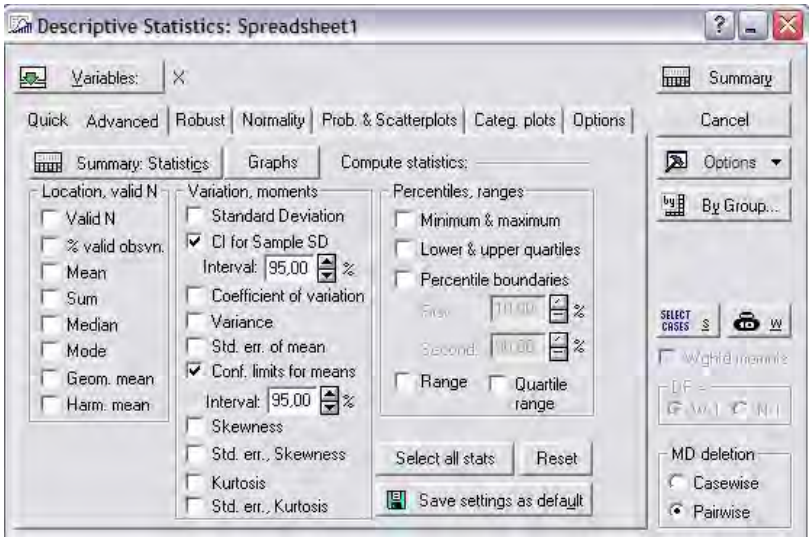


Рис. 2.15 – Задание параметров расчета доверительных интервалов

Результаты вычислений будут представлены в отдельном окне (см. рис. 2.16). Первый столбец содержит имя исходной переменной, четыре других – левую и правую границы доверительного интервала.

	Descriptive Statistics (Spreadsheet1)			
Variable	Confidence -95,000%	Confidence 95,000	Confidence SD -95,000%	Confidence SD +95,000%
X	60,30047	61,40862	1,725447	2,524728

Рис. 2.16 – Результаты вычислений

Задания для самостоятельной работы

Задание 1. В итоге многократных измерений некоторой физической величины одним прибором получены результаты, представленные в таблице. В пакетах *Excel* и *Statistica* необходимо по этим данным:

1) построить дискретный вариационный ряд, провести интервальную обработку;

2) построить полигон частот и гистограмму относительных частот, а также кумулянту

3) вычислить выборочное среднее, моду, медиану распределения, выборочную и исправленные дисперсии, стандартное отклонение, коэффициент вариации, асимметрию и эксцесс, стандартную ошибку среднего. Сделать вывод.

4) определить, сколько нужно иметь наблюдений, чтобы точность определения математичекого ожидания не превышала 0,2.

Вариант 1

11,4	13,89	11,51	12,33	11,29	10,14	9,8	11,31	11,69	10,58
8,91	10,8	10,66	12,38	13,8	11,79	9,94	13,01	10,98	11,11
13,57	10,52	12,16	8,79	10,75	10,74	11,02	12,04	10,19	9,23
11,78	9,94	11,06	12,48	11,49	10,14	11,64	11,63	10,27	10,47
11,24	9,16	9,94	9,42	10,54	12,85	10,68	10,51	12,87	10,14
10,95	10,64	10,25	12,56	10,96					

Вариант 2.

14,8	1,8	16,8	15,1	3,2	-4,6	10,5	10,6	16	11,2
6,9	7,0	14,2	15,7	9,5	2,7	0,4	9,4	1	15,8
14,4	6,4	1,3	11,8	6,4	11,9	21,7	1,2	6,2	1,6
1,9	3,5	4,3	0,3	-2,2	7,8	-0,9	15,4	5,3	15,6
5,2	14,3	11,3	12	7,6	4,7	12,3	4	8,2	12,3

Вариант 3.

13,4	6,0	5,4	12,5	6,3	6,7	0,4	-1,0	11,2	19,3
14,9	13,4	1,3	18,1	0,5	7,7	6,0	10,2	8,3	11,6
5,9	14,2	2,3	6,9	17,8	3,5	2,2	8,4	14,5	4,8
3,1	10,9	7,6	6,6	5,1	-0,7	-9,8	4,1	17,5	4,2
7,3	0,8	14,9	9,7	1,6	7,0	-4,2	-9,2	-4,5	-5,0

Вариант 4

4,5	-0,6	9,6	10,5	15,2	9,3	4,5	9,2	11,9	17,1
4,6	18,3	12,6	4,7	12,9	13,1	14,4	26,3	7,6	7,5
6,8	12,2	10,4	2,6	7,7	12,5	7,2	17,9	11,3	10,3
11,9	8,6	15,6	-0,5	11,1	3	9,7	-1,1	12	13
4,1	13,1	9,3	17,8	6,5	14,3	3,6	17,6	9,3	13,3

Вариант 5

-6,2	1,4	-0,1	-1,0	-3,6	-4,5	6,5	-2,8	0,4	2,1
-11,4	12,3	-2,1	-0,2	-4,8	1,6	3,5	1,7	9,3	-0,5
6,9	-1,1	-1,8	-0,2	-3,7	0,0	2,1	4,5	-0,7	-5,9
3,2	1,4	2,4	6,2	-0,9	6,4	0,6	-4,5	6,8	8,9
-6,9	1,7	3,1	5,1	-2,4	-0,1	-6,0	4,3	-3,4	6,7
0,4	-3,7	8	1,7						

Вариант 6

24,5	16,0	-2,0	14,8	10,0	6,9	8,3	-5,9	14,0	5,3
0,6	14,5	-3,7	0,6	2,1	-12,3	5,9	22,1	20,1	10,3
9,2	15,8	-1,8	1,3	11,2	2,7	9,2	7,6	-1,4	-3,5
27,7	0,9	8,0	8,9	-7,2	5,7	13,5	6,9	-0,3	11,8
21,1	11,6	4,4	-1,9	9,9	14,0	0,9	7,2	21,6	9,7

Вариант 7

43	46	45	43	44	45	47	43	44	46
45	44	42	45	47	44	46	48	46	43
46	43	44	47	45	46	42	44	44	46
47	45	46	46	48	45	45	43	45	47
45	44	45	42	48	48	47	46	46	42

Вариант 8.

7,69	6,19	6,64	5,01	0,55	2,86	3,76	3,99	3,49	3,62
5,34	5,62	8,63	4,62	3,58	4,87	5,37	7,83	7,52	5,52
5,64	3,01	4,66	6,91	7,86	8,55	6,54	4,91	8,84	4,82
3,63	5,08	4,68	6,67	6,03	1,08	2,51	6,71	6,58	3,56
6,94	1,19	5,78	6,04	5,81	6,83	7,00	6,11	6,60	7,39

Вариант 9

1	1,1	1	0,8	0,6	0,6	0,6	0,7	0,7	0,7
1	0,7	0,7	0,6	0,5	0,7	0,7	0,8	0,8	0,9
0,7	1	0,9	0,6	0,7	0,8	0,8	0,8	0,7	0,7
1	0,9	1	1,1	0,8	0,8	0,7	0,6	0,6	0,6
0,9	0,8	0,7	0,7	0,8	0,7	0,8	0,9	0,8	0,8

Вариант 10

20	19	22	24	21	18	23	17	20	16
15	23	21	24	21	18	23	21	19	20
24	21	20	18	17	22	20	16	22	18
20	17	21	17	19	20	20	21	18	22
23	21	25	22	20	19	21	24	23	21
19	22	21	19	20	23	25	25	21	21

Вариант 11

13	10	12	10	12	12	7	9	10	9
11	10	13	9	8	11	6	9	8	11
10	8	10	8	12	10	11	7	12	11
9	8	9	11	9	7	9	10	11	11
10	8	12	9	9	11	15	11	9	10
10	9	10	12	13	11	15	10	7	13

Вариант 12

1	1,1	0,95	0,75	0,58	0,59	0,56	0,69	0,71	0,65
1,02	0,66	0,68	0,57	0,54	0,7	0,72	0,82	0,83	0,91
0,65	0,95	0,85	0,55	0,74	0,76	0,77	0,82	0,66	0,69
0,99	0,88	1,02	1,05	0,81	0,79	0,72	0,61	0,58	0,59
0,91	0,81	0,71	0,73	0,84	0,69	0,82	0,9	0,8	0,77

Лабораторная работа № 3. Статистическая проверка непараметрических гипотез. Критерии согласия.

Необходимые теоретические сведения для выполнения лабораторной работы находятся в теоретическом разделе – тема 3.1.

Контрольный пример 3.1. На основании данных о массе тела студентов (Лабораторная работа № 2, Контрольный пример 2.1), проверить гипотезу о нормальном распределении исследуемой случайной величины, используя приближенную проверку на нормальность, а также с помощью критериев Пирсона (в пакетах *Excel* и *Statistica*), Романовского и Колмогорова (в пакете *Excel*). Принять $\alpha = 0{,}05$.

Решение.

1. *Приближенная проверка с использованием* $\sigma_в$.

Искомая выборка (вариационный ряд):

57	57	58	58	58	58	58	58	59	59	59
59	59	59	59	60	60	60	60	60	60	60
60	60	60	61	61	61	61	61	61	61	61
61	62	62	62	62	62	62	62	62	63	63
63	63	63	63	63	64	64	64	64	65	65

Выборочные числовые характеристики, вычисленные в лабораторной работе № 2: $\bar{x} = 60{,}855; \sigma = 2{,}031$.

Проведем приближенную проверку с использованием оценки σ.

Вычислим значения:

$$0{,}3 \cdot \sigma = 0{,}3 \cdot 2{,}031 = 0{,}6093 \qquad 0{,}7 \cdot \sigma = 0{,}7 \cdot 2{,}031 = 1{,}4217$$
$$0{,}3 \cdot \sigma = 0{,}3 \cdot 2{,}031 = 2{,}2341 \qquad 3 \cdot \sigma = 3 \cdot 2{,}031 = 6{,}093$$

Вычислим границы интервалов:

$$\bar{x} \pm 0{,}3\sigma = 60{,}855 \pm 0{,}6093 = (60{,}3;\ 61{,}5);$$
$$\bar{x} \pm 0{,}7\sigma = 60{,}855 \pm 1{,}4217 = (59{,}4;\ 62{,}3);$$
$$\bar{x} \pm 1{,}1\sigma = 60{,}855 \pm 2{,}2341 = (58{,}6;\ 63{,}1);$$
$$\bar{x} \pm 3\sigma = 60{,}855 \pm 6{,}093 = (54{,}8;\ 66{,}95);$$

Подсчитаем число значений (из общей совокупности), попавших в вычисленные интервалы: $n_1 = 9; n_2 = 27; n_3 = 41; n_4 = 55$.

Вычисляем относительные частоты:

$$w_1 = \frac{9}{55} \approx 0{,}2; \qquad w_2 = \frac{27}{55} \approx 0{,}5; \qquad w_3 = \frac{41}{55} \approx 0{,}75; \qquad w_4 = \frac{55}{55} = 1.$$

Убеждаемся, что во втором, третьем и четвертом интервалах содержится количество случайных чисел не менее рекомендуемого. А в первом – количество чисел близко к рекомендуемому.

Данное эмпирическое распределение, скорее всего, подчиняется нормальному закону распределения, но нужно провести проверку, используя более точные критерии.

Проверим гипотезу о нормальном распределении с помощью критерия Пирсона.

В пакете *Excel.*

Введем исходные данные и оформим их так, как показано на рис. 3.1 (интервальный ряд был получен в лабораторной работе № 2 – табл. 2.2). Вычислим выборочные характеристики, используя стандартные функции пакета. В ячейке G9 найдем общее число наблюдений, просуммировав фактические частоты.

	A	B	C	D	E	F	G
1	**Наблюдения**		**Числовые характеристики**		**xi-1**	**xi**	**Эмпирические частоты**
2	64		**максимум**	**65**	**57**	**58,15**	**8**
3	57		**минимум**	**57**	**58,15**	**59,3**	**7**
4	63		**среднее**	**60,8545**	**59,3**	**60,45**	**10**
5	62		**станд. откл**	**2,0496**	**60,45**	**61,6**	**9**
6	58		**асимметрия**	**0,0963**	**61,6**	**62,75**	**8**
7	61		**эксцесс**	**-0,7437**	**62,75**	**63,9**	**7**
8	63				**63,9**	**65,05**	**6**
9	60				**Всего наблюдений**		**55**
10	60						

Рис. 3.1 – Вычисление выборочных характеристик исходной выборки

В ячейку H2 внесем формулу для вычисления значения функции нормального распределения $F(x_1 = 58{,}15)$. В *Excel* эту величину можно вычислить, воспользовавшись функцией НОРМ.РАСП (рис. 3.2).

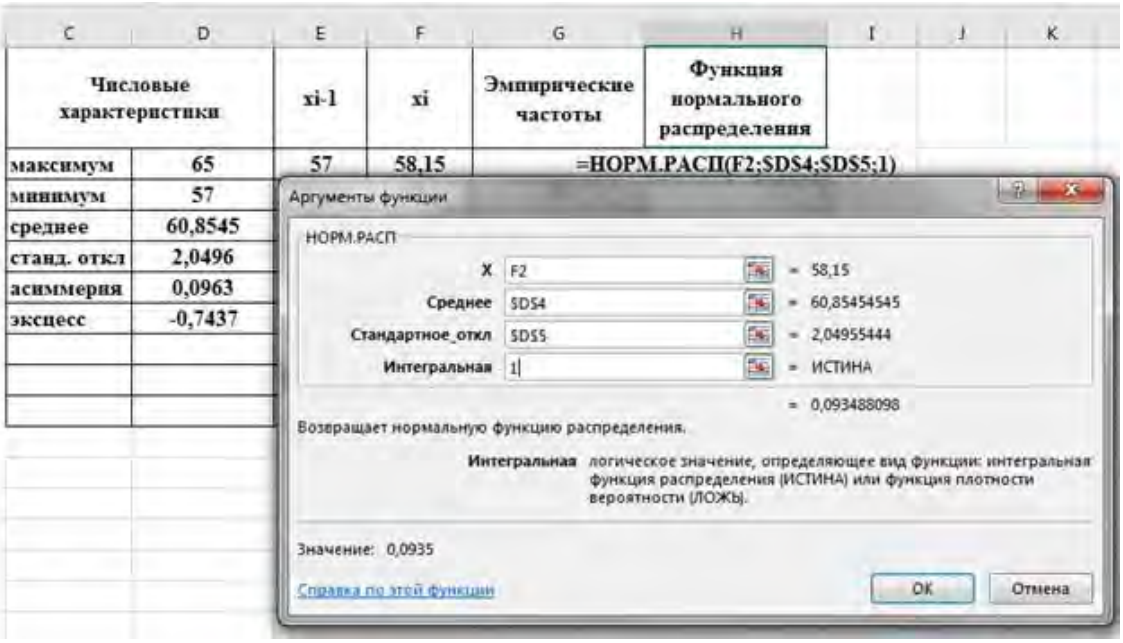


Рис. 3.2 – Диалоговое окно функции НОРМРАСП с заполненными полями ввода

В поле *X* введен адрес ячейки, в которой находится граница первого интервала группировки.

В поле *Среднее* введен адрес ячейки, в которой находится среднее значение выборки.

В поле *Стандартное_откл* введен адрес ячейки, в которой находится значение стандартного отклонения выборки.

В поле *Интегральная* введена единица 1. Единица в поле Интегральная означает вычисление функции распределения *F(x)*.

Далее с помощью маркера автозаполнения протянем эту формулу до ячейки I8.

В ячейку I2 введем формулу =H2, в ячейку I8 – формулу =1–H7, а в ячейку I3 – формулу =H3–H2, протянув ее до ячейки I7 (рис. 3.3).

В результате этих действий в диапазоне H2:H8 появятся значения теоретических вероятностей $p_1, p_2, \ldots, p_7$, причем $p_1 + p_2 + \ldots + p_7 = 1$.

В ячейку J2 введем формулу =I2*$G$9 и протянем ее до ячейки J8. С помощью этой формулы вычисляются теоретические частоты n_i'. Их сумма равна объему выборки $n = 55$. На этом заканчивается первый этап проверки (рис. 3.3).

E	F	G	H	I	J
xi-1	**xi**	**Эмпирические частоты**	**Функция нормального распределения**	**Теоретические вероятности**	**Теоретические частоты**
57	58,15	8	0,0935	0,0935	5,1418
58,15	59,3	7	0,2241	0,1306	7,1827
59,3	60,45	10	0,4218	0,1977	10,8725
60,45	61,6	9	0,6420	0,2202	12,1110
61,6	62,75	8	0,8225	0,1805	9,9277
62,75	63,9	7	0,9313	0,1089	5,9884
63,9	65,05	6	0,9797	0,0687	3,7758
Всего наблюдений		55		1	55

Рис. 3.3 – Первый этап проверки гипотезы по критерию хи-квадрат

Результаты заключительного этапа проверки приведены в диапазоне L1:N13 (рис. 3.4). В столбце U в ячейке N11 будет вычисляться наблюдаемое значение критерия.

В диапазоне L2:L8 находятся эмпирические частоты n_i в диапазоне M2:M8 – теоретические частоты n_i'.

В ячейку N2 введена формула =(L2-M2)^2/M2, реализующая вычисления по формуле $\frac{(n_i - n_i')^2}{n_i'}$. Размножим эту формулу в диапазоне ячеек N3:N8. В ячейке N11 получим сумму содержимого ячеек N2:N8 по формуле =СУММ(N2:N8).

Критическое значение статистики U, которая имеет распределение с $k = 7 - 3 = 4$ степенями свободы, определяется при помощи функции ХИ2.ОБР.ПХ (0,05; 4).

Расчетное значение $\chi^2_{\text{набл}} = 4{,}3179$ статистики U меньше ее критического значения $\chi^2_{\text{набл}} = 9{,}4877$ (рис. 3.4), поэтому можно сказать, что проверяемая гипотеза, состоящая в том, что генеральная совокупность подчиняется нормальному закону распределения, не противоречит данным эксперимента.

Согласно правилу Романовского $c = \frac{|4{,}3179 - 4|}{\sqrt{2 \cdot 4}} \approx 0{,}049 < 3$.

Так как $c < 3$. расхождение между гипотетическим и эмпирическим распределениями следует считать случайным. Основная гипотеза H_0 принимается, т.е. генеральное распределение считается нормальным.

E	F	G	H	I	J	K	L	M	N
xi-1	xi	Эмпирические частоты	Функция нормального распределения	Теоретические вероятности	Теоретические частоты		Эмп. частоты	Теорет. частоты	U
57	58,15	8	0,0935	0,0935	5,1418		8	5,1418	1,5887
58,15	59,3	7	0,2241	0,1306	7,1827		7	7,1827	0,0046
59,3	60,45	10	0,4218	0,1977	10,8725		10	10,8725	0,0700
60,45	61,6	9	0,6420	0,2202	12,1110		9	12,1110	0,7991
61,6	62,75	8	0,8225	0,1805	9,9277		8	9,9277	0,3743
62,75	63,9	7	0,9313	0,1089	5,9884		7	5,9884	0,1709
63,9	65,05	6	0,9797	0,0687	3,7758		6	3,7758	1,3102
Всего наблюдений		55		1	55		55	55	
							Наблюдаемое значение Хи-квадрат		
									4,3179
							Критическое значение Хи-квадрат		
									9,4877

Рис. 3.4 – Таблица с окончательными результатами вычисления статистики

2. *Проверка истинности гипотезы H_0 по критерию Колмогорова*.

Скопируем данные наблюдений на чистый лист *Excel* и построим интервальный вариационный ряд (рис. 3.5).

Рассчитаем выборочную среднюю и «исправленное» стандартное отклонение, используя функции СРЗНАЧ(A2:A56) и СТАНДОТКЛОН.В(A2:A56) – в ячейках D4 и D5.

Вычислим статистику λ в граничных точках интервального ряда (см. рис. 3.5)

A	B	C	D	E	F	G	H	I	J	K	L
Наблюдения		Числовые характеристики		xi-1	xi	Частоты	Относительные частоты	Накопленные частоты (F*(x))	(xi-xs)/s	F(xi)	|F*(x)-F(x)|
64		максимум	65	менее 57		0	0	0	-29,6916	0	0,0000
57		минимум	57	57	58,15	8	0,1455	0,1455	-1,31958	0,093488	0,0520
63		среднее xs	60,85455	58,15	59,3	7	0,1273	0,2727	-0,75848	0,224082	0,0486
62		станд_откл s	2,049554	59,3	60,45	10	0,1818	0,4545	-0,19738	0,421764	0,0328
58				60,45	61,6	9	0,1636	0,6182	0,363715	0,641965	0,0238
61				61,6	62,75	8	0,1455	0,7636	0,924813	0,822468	0,0588
63				62,75	63,9	7	0,1273	0,8909	1,485911	0,931349	0,0404
60				63,9	65,05	6	0,1091	1,0000	2,047008	0,979671	0,0203
60				Всего наблюдений		55	1				0,0588

Рис. 3.5 – Расчеты для критерия Колмогорова

В столбцах H и I рассчитаем относительные и накопленные частоты (см. лабораторную работу № 2). Накопленные частоты (точки кумулятивной кри-

вой) – левые концы «ступенек» эмпирической функции распределения, т. е. $F^*(x)$.

Значения $F(x_i)$ вычисляются с учетом того, что была выдвинута гипотеза о нормальном распределении. В пакете *Excel* значения $F(x_i)$ можно вычислить с помощью функции НОРМ.СТ.РАСП.

Из последнего столбца таблицы ясно, что $\max_i \left|F^*(x_i) - F(x_i)\right| = 0{,}0588$. Тогда значение статистики Колмогорова: $\lambda = \sqrt{55} \cdot 0{,}0588 \approx 0{,}436$.

По заданному уровню значимости $\alpha = 0{,}05$ определим границу критической области (табл. 3.1 из теоретического раздела) $\lambda_{крит} = 1{,}3581$. Поскольку $\lambda = 0{,}436 < \lambda_{крит} = 1{,}3581$ то основная гипотеза принимается, т. е. генеральное распределение считается нормальным.

3. Проверим гипотезу о нормальном распределении случайной величины в пакете *Statistica*.

После ввода исходных данных будем использовать процедуру *Distribution Fitting* (подбор распределения), как показано на рис. 3.6.

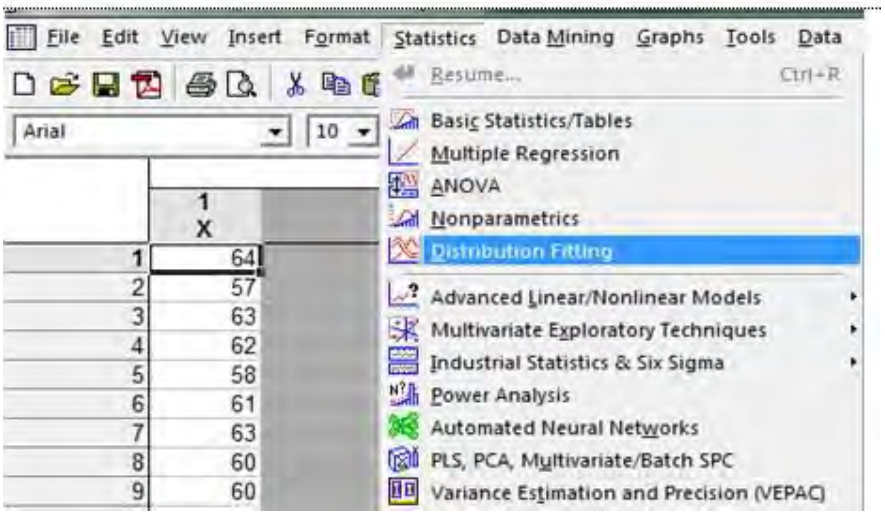


Рис. 3.6 – Вызов процедуры *Distribution Fitting*

На вкладке *Quick* выбираем непрерывные распределения (*Continuos Distributions*) – *Normal* (если выдвигается гипотеза об экспоненциальном распределении – *Exponential*; если о равномерном – *Rectangular*).

Зададим диапазон исходных данных, нажав на кнопку *Variable* и выбрав там X. Далее нажмем кнопку *OK*.

Во вкладке *Parameters* установим количество интервалов разбиения (*Number of categories*), равное 7 (рис. 3.7), а также минимальное (*Lower limit*) и максимальное (*Upper limit*) значения выборки.

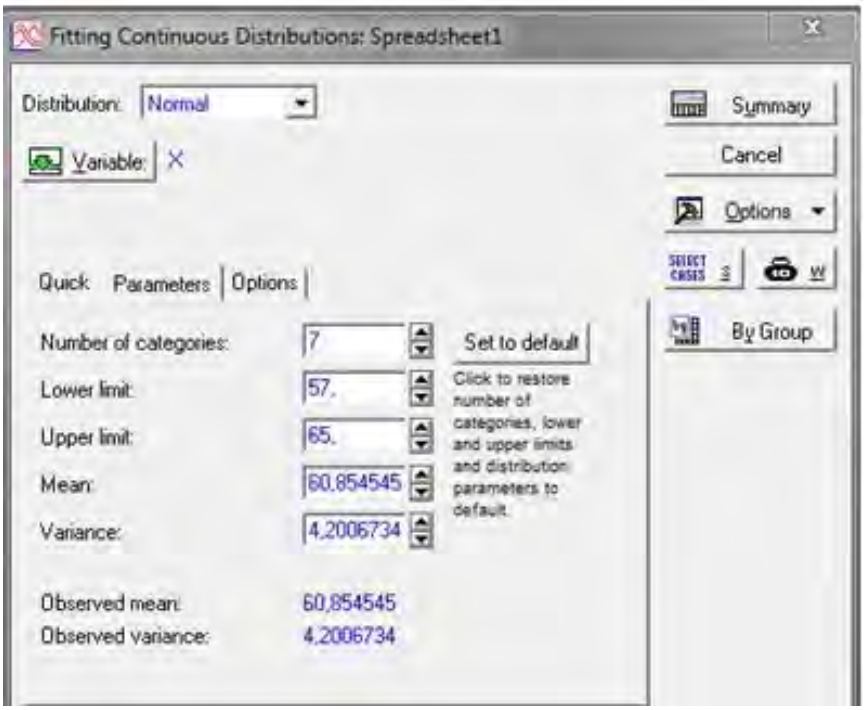


Рис. 3.7 – Вкладка *Parameters*

Нажав кнопку *Summary*, получаем таблицу частот (рис. 3.8).

Variable: X, Distribution: Normal (Spreadsheet1)
Chi-Square = 3,73766, df = 3 (adjusted) , p = 0,29122

Upper Boundary	Observed Frequency	Cumulative Observed	Percent Observed	Cumul. % Observed	Expected Frequency	Cumulative Expected	Percent Expected	Cumul. % Expected	Observed-Expected
<= 58,14286	8	8	14,54545	14,5455	5,10990	5,10990	9,29073	9,2907	2,89010
59,28571	7	15	12,72727	27,2727	7,10020	12,21010	12,90945	22,2002	-0,10020
60,42857	10	25	18,18182	45,4545	10,76219	22,97229	19,56761	41,7678	-0,76219
61,57143	9	34	16,36364	61,8182	12,04876	35,02104	21,90683	63,6746	-3,04876
62,71429	8	42	14,54545	76,3636	9,96340	44,98445	18,11528	81,7899	-1,96340
63,85714	7	49	12,72727	89,0909	6,08523	51,06968	11,06405	92,8540	0,91477
< Infinity	6	55	10,90909	100,0000	3,93032	55,00000	7,14604	100,0000	2,06968

Рис. 3.8 – Таблица частот

Если гипотеза верна, вероятность получить 3,73766 или больше равна 0,29122 (больше 0,05 – уровня значимости) – достаточна, чтобы поверить в нормальность распределения исходных данных. Следовательно, гипотезу о нормальном распределении случайной величины принимаем.

Вывод. На основе критериев Пирсона и Колмогорова, которые дали аналогичные результаты, гипотезу о нормальном распределении генеральной случайной величины следует принять.

Задание для самостоятельной работы

Задание 1. Используя данные своего варианта из лабораторной работы № 2:

– провести приближенную проверку на нормальность;

– проверить, согласуются ли выборочные данные с гипотезой о нормальном распределении с помощью критериев Пирсона (в пакетах Excel, Statistica), Романовского и Колмогорова (в пакете Excel). Принять $\alpha = 0{,}05$.

Лабораторная работа № 4. Проверка гипотез о параметрах распределения

Необходимые теоретические сведения для выполнения лабораторной работы находятся в теоретическом разделе – тема 3.2.

Контрольный пример 4.1. Согласно технической документации, среднее время срабатывания взрывателя ручной гранаты равно 4 секунды. При проверке 12 взрывателей зафиксированы следующие значения времени срабатывания:

4,02; 3,92; 4,07; 4,18; 4,17; 4,23; 3,83; 4,03; 4,16; 3,94; 3,98; 3,88.

1. Проверить на уровне значимости $\alpha = 0{,}05$ гипотезу $H_0 : a = 4$ о том, что среднее время срабатывания взрывателя равно 4 секунды. Принять в качестве $H_0 : a \neq 4$. Задание выполнить в пакетах *Statistica* и *Excel*.

2. Проверить на уровне значимости $\alpha = 0{,}05$ гипотезу о том, что стандартное отклонение σ времени X срабатывания ручной гранаты равно 0,1 с. Задание выполнить в пакетах *Excel и Mathcad.*

Решение.

1. Проверим гипотезу в пакете *Excel*.

В диапазон A1:A12 листа *Excel* введем исходные данные наблюдения (рис. 4.1).

D9 | =СТЬЮДЕНТ.ОБР(1-0,05;11)

	A	B	C	D	E
1	4,02		Выб. Среднее =	4,034167	
2	3,92		S =	0,129927	
3	4,07		t =	0,910948	
4	4,18		**Критическая точка для двусторонней критич. области**		
5	4,17		t(0.025, 11) =	2,200985	
6	4,23		**Нахождение критической точки с помощью функции**		
7	3,83		СТЬЮДЕНТ.ОБР(1-0,025;11)=	2,200985	
8	4,03		**Критическая точка для односторонней критич. области**		
9	4,16		t_прав_крит(0.05;11)=	1,795885	
10	3,94				
11	3,98				
12	3,88				

Рис. 4.1– Проверка гипотезы о среднем времени срабатывания взрывателя гранаты

В ячейки C1:C5 введем информационные метки: Выб. среднее =. S = , t = , t(0,025; 11) =

В ячейку D1 введем формулу = СРЗНАЧ(A1:A12) и нажмем клавишу *Enter*. В ячейке D1 появится выборочное среднее времени срабатывания взрывателя. Ручная граната непригодна для боевого использования, если среднее время срабатывания взрывателя слишком мало или слишком велико (противник успеет бросить гранату обратно). Поэтому в качестве H_1 используется гипотеза $a \neq a_0$. Такой альтернативе соответствует двусторонняя критическая область $K_{\text{кр}}(0{,}05) = \left(|T_{11}| \geq t(0{,}025;11)\right)$, где $t(0{,}025;11)$ – критическое значение распределения Стьюдента с 11 степенями свободы.

В ячейку D2 введем формулу = СТАНДОТКЛ.В(A1:A12) и нажмем клавишу *Enter*. В ячейке появится выборочное стандартное отклонение $S = 0{,}129927$ времени срабатывания взрывателя.

В ячейку D3 введем формулу =(D1–4)*КОРЕНЬ(12)/D2 и нажмем клавишу *Enter*.

В ячейку D5 введем формулу =СТЬЮДЕНТ.ОБР.2Х(0,05;11) и нажмем клавишу *Enter*. В ячейке появится критическое значение $t_{\text{кр}}(0{,}025;11) = 2{,}2$ порядка 0,025 распределения Стьюдента с 11 степенями свободы.

Полученный результат $\left(|t| < t\left(\frac{\alpha}{2}, n-1\right)\right)$ свидетельствует о том, что гипотеза $H_0 : a = 4$ не противоречит данным эксперимента.

Замечание 4.1. Критическое значение можно найти с помощью стандартной функции =СТЬЮДЕНТ.ОБР(1–0,025;11) – результат будет аналогичным.

В пакете *Statistica.*

Введем исходные данные.

Для проверки гипотезы будем использовать процедуру Basic *Statistics/Tables*, которая находится в меню *Statistics*.

Выбираем *t-test*, *single sample*.

Зададим диапазон исходных данных, нажав на кнопку *Variables*. Нажимаем кнопку *OK*. Далее на вкладке *Advanced* зададим исходные данные задачи (рис. 4.2).

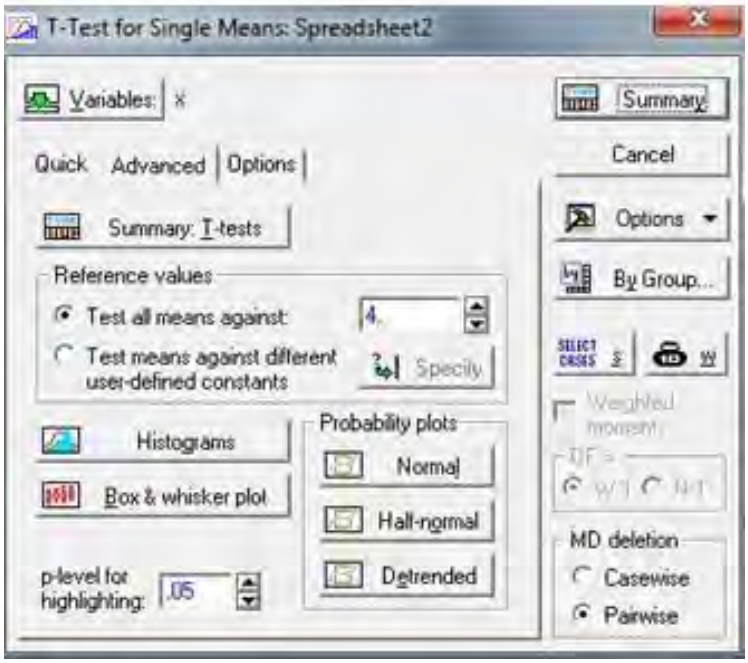


Рис. 4.2 – Исходные данные задачи на вкладке *Advanced*

Нажав кнопку *Summary*, получаем таблицу (рис. 4.3).

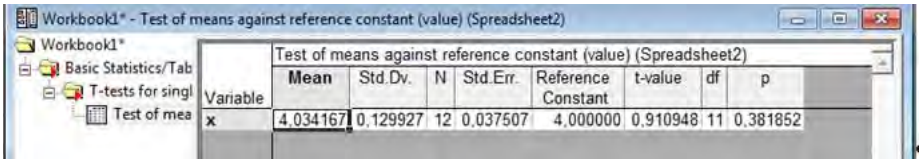


Рис. 4.3 – Проверка гипотезы о математическом ожидании при неизвестной дисперсии в пакете *Statistica*

Так как $p = 0{,}381852 > 0{,}05$, то приходим к выводу, что проверяемая гипотеза не противоречит данным эксперимента.

Проверяемая гипотеза эквивалентна гипотезе $H_0 : \sigma = 0.1^2 = 0.01$. Увеличение разброса времени срабатывания взрывателя влечёт за собой возрастание опасности самоподрыва и опасности использования гранаты противником. В связи с этим в качестве конкурирующей надо выбрать гипотезу $H_1 : \sigma > \sigma_0$.

На рис. 4.4 представлено решение данной задачи в пакете Excel.

C	D	E	F	G	H
Выб. Среднее =	4,034167		$\chi^2_{набл}$ =	18,56917	
S =	0,129927				
t =	0,910948		$\chi^2_{кр}$ =	19,67514	
t(0.025, 1) =	2,200985				

Рис. 4.4 – Проверка гипотезы о значении дисперсии

В ячейку G1 введена формула =D2^2*11/0,01, в ячейку G3 – функция =ХИ2.ОБР.ПХ(0,05;11).

На рисунке 4.5 представлено решение данной задачи в пакете Mathcad.

Исходные данные

$x := (4.02 \;\; 3.92 \;\; 4.07 \;\; 4.18 \;\; 4.17 \;\; 4.23 \;\; 3.83 \;\; 4.03 \;\; 4.16 \;\; 3.94 \;\; 3.98 \;\; 3.88)^T$

$s2 := Var(x)$ $\quad s2 = 0.017$ $\quad \sigma 0 := 0.01$

Объём выборки $\quad n := rows(x) \quad n = 12$

Число степеней свободы $\quad k := n - 1 \quad k = 11$

Расчётное значение критерия

$\chi 2 := \frac{n - 1}{\sigma 0} \cdot s2 \quad \chi 2 = 18.569$

Граница критической области $\quad qchisq(0.95, 11) = 19.675$

Рис. 4.5 – Проверка гипотезы о значении дисперсии в пакете Mathcad.

Сравнивая расчетное значение статистики χ^2 с ее критическим значением порядка 0,05, приходим к выводу, что проверяемая гипотеза не противоречит данным наблюдения.

Если рассматривается гипотеза $\sigma^2 \neq \sigma_0^2$:

Расчётное значение критерия $\quad \chi 2 := \frac{n - 1}{\sigma 0} \cdot s2 \quad \chi 2 = 18.569$

Для двусторонней критической области

$\chi\_п\_кр := qchisq\left(1 - \frac{0.05}{2}, 11\right) = 21.92$

$\chi\_л\_кр := qchisq\left(\frac{0.05}{2}, 11\right) = 3.816$

Так как χ_л_кр<χ2<χ_п_кр - нет оснований отклонить нулевую гипотезу

Рис. 4.6 – Проверка гипотезы о значении дисперсии в пакете Mathcad при $H_1 : \sigma^2 \neq \sigma_0^2$.

Контрольный пример 4.2. Проверить гипотезу о равенстве математических ожиданий $M(X)=a_1$ и $M(Y)=a_2$ двух независимых нормально распределённых случайных величин (малые выборки) при уровне значимости $\alpha=0{,}05$.

X	1.08	1.1	1.12	1.14	1.15	1.25	1.36	1.38	1.4	1.42
Y	1.11	1.12	1.18	1.22	1.33	1.35	1.36	1.38		

Предварительно проверить гипотезу о равенстве дисперсий.

Задания выполнить в пакетах Statistica и Excel

Решение. При проверке гипотезы о равенстве дисперсий в пакете *Excel* используются статистическая процедура «*Двухвыборочный F-тест для дисперсий*» и встроенная статистическая функция F.ТЕСТ, которая рассчитывает *p*-значение.

Введем выборочные данные в столбцы A и B листа *MS Excel* (рис. 4.7 – диапазон A1:B11). С помощью функции ДИСП.В рассчитаем «исправленные» дисперсии s_x^2 и s_y^2 по этим выборкам – результаты в ячейках E1 и E2 (рис. 4.7). Так как s_x^2 больше, то именно диапазон выборочных значений X должен выступать в качестве первой переменной при использовании инструмента анализа *Двухвыборочный F-тест для дисперсий*.

1) Выбираем раздел меню *Данные – Анализ данных – Двухвыборочный F-тест для дисперсии* и нажмём клавишу *Enter***.**

2) В поле *Интервал переменной 1* введём диапазон A1:A11, а в поле *Интервал переменной 2* – диапазон B1:B8.

3) В поле Альфа введём число 0,025, равное половине заданного уровня значимости $\alpha=0{,}025$ (это обусловливается тем, что в данном примере рассматривается двусторонняя альтернатива $H_1:\sigma_1^2\neq\sigma_2^2$).

4) В группе переключателей *Параметры вывода* выберем переключатель *Выходной Интервал.* В открывшемся справа от этого переключателя поле введём ссылку на ячейку D4, в которой расположится левый верхний угол таблицы результатов решения. Щелкнем на кнопке OK.

5) На экране в диапазоне D4:F13 появится таблица результатов (рис. 4.7).

E15 =F.TECT(A2:A11;B2:B8)

	A	B	C	D	E	F
1	**X**	**Y**		**"Исправленная" дисперсия X**	0,018521111	
2	1,09	1,11		**"Исправленная" дисперсия Y**	0,012890476	
3	1,1	1,12				
4	1,12	1,18		Двухвыборочный F-тест для дисперсии		
5	1,14	1,22				
6	1,15	1,33			*X*	*Y*
7	1,25	1,36		Среднее	1,241	1,24286
8	1,36	1,38		Дисперсия	0,01852	0,01289
9	1,38			Наблюдения	10	7
10	1,4			df	9	6
11	1,42			F	1,43681	
12				P(F<=f) одностороннее	0,33983	
13				F критическое одностороннее	5,52341	
14						
15				Проверка с помощью функции F.TECT	0,67966	

Рис. 4.7 – Исходные данные и результаты решения первой части контрольного примера 4.2

Здесь символом F обозначено наблюдаемое значение статистики Фишера. Символ $P(F \le f)$ обозначает статистическую значимость $p = P(F(k_1, k_2)) \ge f$ – если $p \le \alpha$, то нулевая гипотеза отвергается, а символ *F критическое одностороннее* – критическое значение $F_{\text{kp}}(\alpha, k_1, k_2)$ порядка α распределения Фишера с $k_1 = n_1 - 1$ и $k_2 = n_2 - 1$ степенями свободы.

Анализ результатов решения свидетельствует о том, что наблюдаемое значение 1,43681 статистики F меньше её критического значения $F_{\text{kp}}(0{,}025; 9; 6) = 5{,}52341$ порядка 0,025. Это означает, что проверяемая гипотеза не противоречит фактическим данным наблюдения и её можно принять. К такому же выводу приводит сравнение значимости $p = 2P = 2 \cdot 0{,}33983 = 0{,}67966$ с заданным уровнем значимости $\alpha = 0{,}05$: гипотезу H_0 можно принять, так как $p > \alpha$.

Замечание 4.2. На рисунке 4.7 в ячейках D15:E15 приведена проверка гипотезы о равенстве дисперсий с помощью функции **F.TECT.** В ячейке E15 приведена значимость $p = 2P(F(0{,}05; 11; 11) > 1{,}109) = 0{,}8668$ (полученное значение совпадает с удвоенным значением числа, находящегося в ячейке E12). Полученный результат $(p > \alpha)$ свидетельствует о том, что гипотеза H_0 не противоречит данным наблюдения.

Замечание 4.3**.** Если бы в качестве альтернативы выступала гипотеза $H_1 : \sigma_1^2 > \sigma_2^2$, в поле **Альфа** надо было ввести число 0,05. При этом процедура выдала бы следующие результаты (рис. 4.8).

Двухвыборочный F-тест для дисперсии		
	X	*Y*
Среднее	1,241	1,24286
Дисперсия	0,01852	0,01289
Наблюдения	10	7
df	9	6
F	1,43681	
P(F<=f) одностороннее	0,33983	
F критическое одностороннее	4,09902	

Рис. 4.8 – Результаты решения контрольного примера 4.2 при $H_1 : \sigma_1^2 > \sigma_2^2$

Т.е. $F_{\text{набл}} = 1{,}43681 < F_{\text{кр}}(0{,}05; 9; 6) = 4{,}09902$. Таким образом, и при альтернативе $H_1 : \sigma_1^2 > \sigma_2^2$ проверяемая гипотеза не противоречит данным наблюдения.

В пакете *Excel* проверяется гипотеза о том, что *разность* между математическими ожиданиями независимых нормально распределённых случайных величин X и Y с одинаковыми неизвестными дисперсиями равна заданному числу δ.

Предварительно было установлено, что выборочные дисперсии оценок различаются незначимо (несущественно).

Для проверки гипотезы $H_0 : a_1 - a_2 - \delta = 0$ в пакете *Excel* используется статистическая процедура *Двухвыборочный t-тест с одинаковыми дисперсиями.*

а) Скопируем на новый рабочий лист диапазон ячеек A1:B11, на котором записаны значения выборок X и Y (рисунок 4.9).

б) Выбираем раздел меню *Данные – Анализ данных – Двухвыборочный t-тест с одинаковыми дисперсиями* и нажмём клавишу *Enter*.

в) Заполним диалоговое окно этой процедуры так, как показано на рис. 4.9 и щёлкнем на кнопке *ОК*.

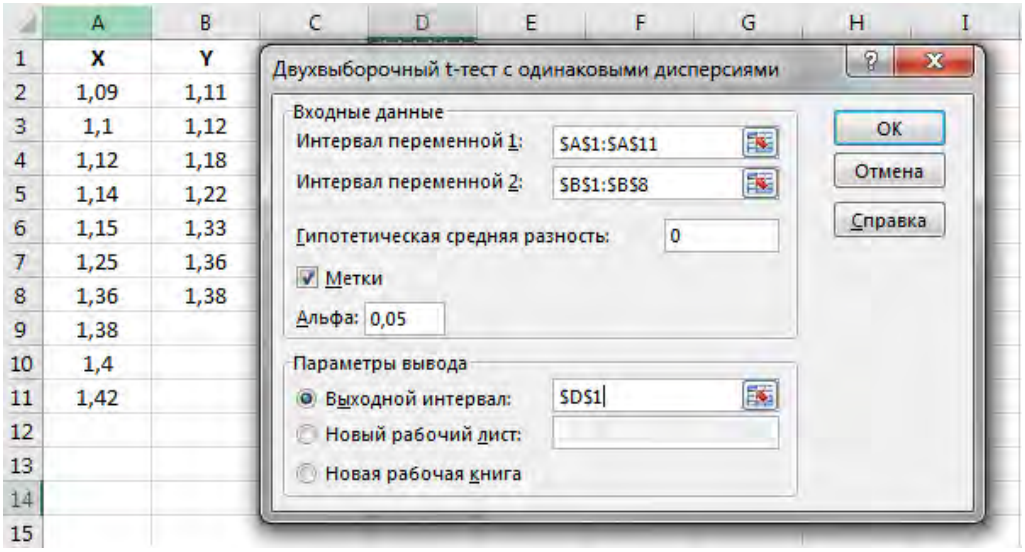


Рис. 4.9 – Диалоговое окно процедуры Двухвыборочный t-тест с одинаковыми дисперсиями

г) На экране в диапазоне D1:F14 появится таблица результатов решения (см. рис. 4.10)

	A	B	C	D	E	F
1	**X**	**Y**		Двухвыборочный t-тест с одинаковыми дисперсиями		
2	1,09	1,11				
3	1,1	1,12			*X*	*Y*
4	1,12	1,18		Среднее	1,241	1,24286
5	1,14	1,22		Дисперсия	0,01852	0,01289
6	1,15	1,33		Наблюдения	10	7
7	1,25	1,36		Объединенная дисперсия	0,01627	
8	1,36	1,38		Гипотетическая разность средних	0	
9	1,38			df	15	
10	1,4			t-статистика	-0,02955	
11	1,42			P(T<=t) одностороннее	0,48841	
12				t критическое одностороннее	1,75305	
13				P(T<=t) двухстороннее	0,97682	
14				t критическое двухстороннее	2,13145	

Рис. 4.10 – Проверка гипотезы о равенстве средних двух нормальных генеральных совокупностей с неизвестными равными дисперсиями

Анализ результатов решения свидетельствует о том, что расчётное значение статистики T находится в области принятия гипотезы (–2,13145; 2,13145). Это означает, что гипотеза о равенстве средних показателей a_1 и a_2 не противоречит фактическим данным наблюдения и, следовательно, её принять (на уровне значимости $\alpha = 0{,}05$). К такому же выводу приводит и сравнение значимости $p = 0{,}97682$ с заданным уровнем значимости $\alpha = 0{,}05$: гипотезу H_0 следует принять, так как $p > \alpha$.

Проверим гипотезу в пакете *Statistica*.

Для проверки гипотезы будем использовать процедуру *Basic Statistics/Tables*, которая находится в меню *Statistics*.

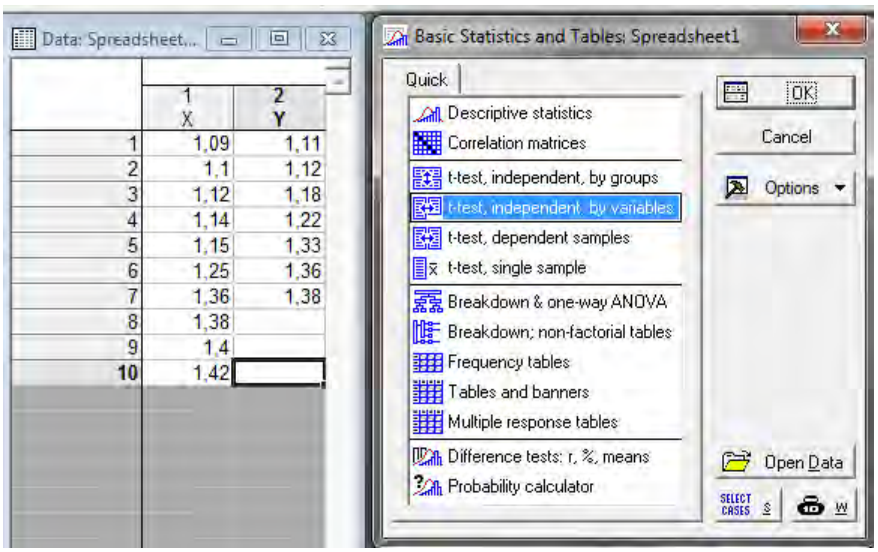


Рис. 4.11 – Исходные данные и выбор модуля
t-test, independent, by variables

Выбираем *t-test, independent, by variables* (рис. 4.11). Зададим диапазон исходных данных, нажав на кнопку *Variables* и выбрав там: *X* и *Y*. Далее нажмём кнопку OK.

Нажав кнопку *Summary*, получаем таблицу (рис.4.12):

T-test for Independent Samples (Spreadsheet1)
Note: Variables were treated as independent samples

Group 1 vs. Group 2	Mean Group 1	Mean Group 2	t-value	df	p	Valid N Group 1	Valid N Group 2	Std.Dev. Group 1	Std.Dev. Group 2	F-ratio Variances	p Variances
X vs. Y	1,241000	1,242857	-0,029546	15	0,976819	10	7	0,136092	0,113536	1,436806	0,679664

Рис. 4.12 – Проверка гипотезы о равенстве средних двух нормальных генеральных совокупностей с неизвестными одинаковыми дисперсиями в пакете *Statistica*

Гипотеза о равенстве математических ожиданий принимается, так как $p = 0{,}976819 > \alpha = 0{,}05$ – для двусторонней критической области, т.к. пакет *Statistica* проводит вычисления при альтернативе $H_1 : a_1 \neq a_2$.

Замечание 4.4. В модуле *t-test, independent, by variables* предполагается уже установленное равенство дисперсий, поэтому число степеней свободы df всегда равно $n_1 + n_2 - 2$. Для контроля выполнения этого условия в столбце *F-ratio variances* приводится вычисленное значение *F*-критерия. Если это значение превышает табличное (или *p Variance* < α), следует воспользоваться другой формулой для проверки гипотезы о равенстве средних, либо воспользоваться непараметрическими критериями сравнения двух выборок.

В данном примере гипотеза о равенстве дисперсий принимается, так как *p Variance* = 0,679664 > 0,05.

*Контрольный пример 3.3***.** При измерении производительности двух агрегатов получены следующие результаты (в кг вещества за час работы):

№ замера	1	2	3	4	5
Агрегат X	14,1	10,1	14,7	13,7	14
Агрегат Y	14	14,5	13,7	12,7	14,1

Можно ли считать, что производительности агрегатов X и Y одинаковы, в предположении, что обе выборки получены из нормально распределённых генеральных совокупностей? Принять $\alpha = 0.1$. Задание выполнить в пакетах Excel и Mathcad.

Предварительно проверить гипотезу о равенстве дисперсий (в пакете Excel).

*Решение***.** Проверяется гипотеза $H_0 : a_1 = a_2$ при альтернативной гипотезе $H_1 : a_1 \neq a_2$.

Работаем в пакете Excel. Проверим гипотезу о равенстве дисперсий:

$$H_0 : \sigma_1 = \sigma_2,\ H_1 : \sigma_1 \neq \sigma_2.$$

Проверку проведём с помощью стандартной функции F.ТЕСТ (см. рис. 4.13):

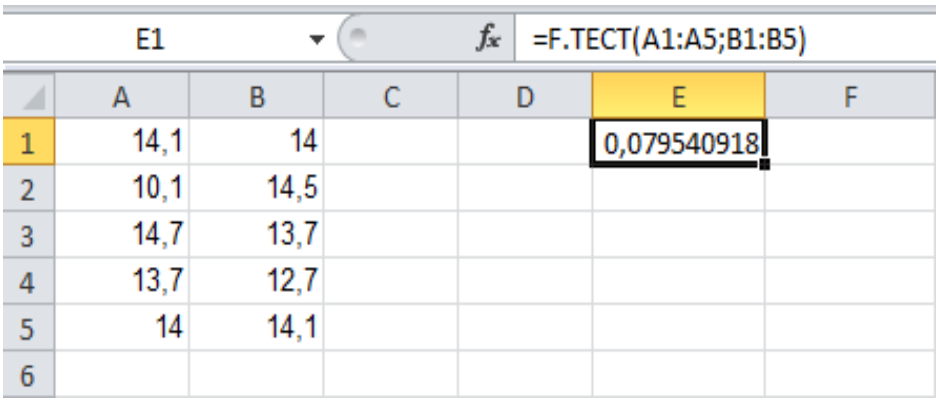


Рис. 4.13 – Проверка гипотезы о равенстве дисперсий
с помощью функции F.ТЕСТ

В ячейке E1 – значимость $p = 0{,}07954$. Так как $p < \alpha = 0.1$ – гипотеза H_1 противоречит опытным данным.

Для проверки гипотезы о равенстве (в пакете Excel – о разности) математических ожиданий двух независимых нормальных случайных величин с различными дисперсиями используется статистическая процедура **Двухвыборочный t-тест с различными дисперсиями** и встроенная статистическая функция **ТТЕСТ**.

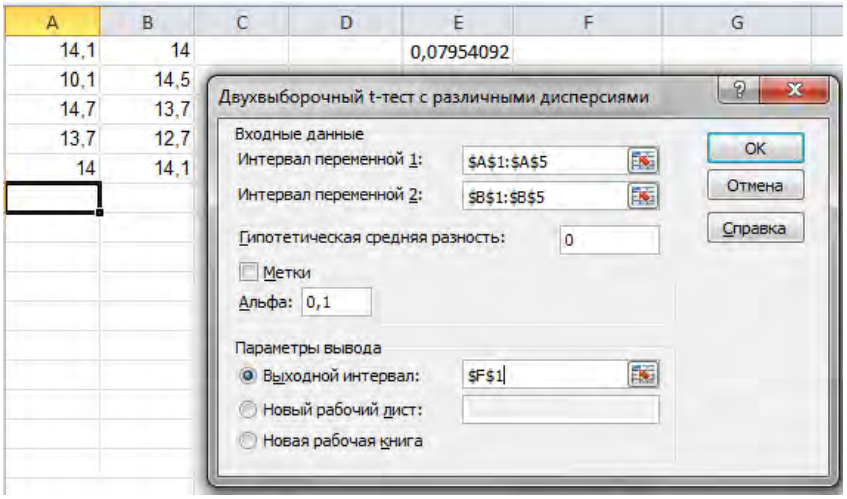


Рис. 4.14. Диалоговое окно процедуры
Двухвыборочный t-тест с различными дисперсиями

Результаты проверки приведены на рисунке 4.15 в диапазоне ячеек F1:H13.

J5 f_x

A	B	C	D	E	F	G	H
14,1	14			0,07954092	Двухвыборочный t-тест с различными дисперсиями		
10,1	14,5						
14,7	13,7					*Переменная 1*	*Переменная 2*
13,7	12,7				Среднее	13,32	13,8
14	14,1				Дисперсия	3,372	0,46
					Наблюдения	5	5
					Гипотетическая разность средних	0	
					df	5	
					t-статистика	-0,548293995	
					P(T<=t) одностороннее	0,303535648	
					t критическое одностороннее	1,475884049	
					P(T<=t) двухстороннее	0,607071296	
					t критическое двухстороннее	2,015048373	

Рис. 4.15 – Проверка гипотезы о равенстве средних двух нормальных генеральных совокупностей с неизвестными различными дисперсиями.

Анализ результатов решения свидетельствует о том, что расчётное значение $t = -0.548$ статистики T находится в области принятия нулевой гипотезы: $|t| < t_{\text{двустор.кр}}(0.1;5) = 2.015$.

Это означает, что гипотеза о равенстве средних принимается – можно считать, что производительность агрегатов X и Y – одинакова.

Проверка гипотезы о равенстве математических ожиданий в пакете Mathcad приведена на рис 4.16. Числовые характеристики выборки рассчитаны ранее в пакете Excel (см рис. 4.15).

Числовые характеристики исходных выборок Объемы выборок

$xcp := 13.32$ $ycp := 13.8$ $n1 := 5$

$dx := 3.372$ $dy := 0.46$ $n2 := 5$

Наблюдаемое значение критерия, рассчитанное по формуле (3.6)

$$T := \frac{xcp - ycp}{\sqrt{\frac{dx}{n1} + \frac{dy}{n2}}} = -0.548$$

Число степеней свободы

$$k := \text{trunc}\left[\frac{\left(\frac{dx}{n1} + \frac{dx}{n2}\right)^2}{\frac{\left(\frac{dx}{n1}\right)^2}{n1 - 1} + \frac{\left(\frac{dx}{n2}\right)}{n1 - 2}}\right] = 5$$

Критическое значение двусторонней критической области

$$tkp := qt\left(1 - \frac{0.1}{2}, k\right) = 2.015$$

$|T| < tkp = 1$ Гипотеза о равенстве математических ожиданий принимается

Рис 4.16 – Проверка гипотезы о равенстве математических ожиданий в пакете Mathcad

Здесь функция trunc(z) – стандартная функция пакета, которая возвращает целую часть от z, удаляя дробную часть.

Контрольный пример 4.4. На двух аналитических весах, в одном и том же порядке, взвешены 10 проб химического вещества и получены следующие результаты взвешиваний (в мг):

x_i	25	30	28	50	20	40	32	36	42	38
y_i	28	31	26	52	24	36	33	35	45	40

При уровне значимости 0,01 установить, значимо или незначимо различаются результаты взвешиваний, в предположении, что они распределены нормально.

Задание выполнить в пакетах *Excel* и *Statistica*.

Принять в качестве альтернативной гипотезу $H_1 : a_1 \neq a_2$.

Решение. В пакете *Excel* проверяется гипотеза $H_0 : a_1 = a_2 = \delta$ о разности математических ожиданий двух коррелированных (зависимых) нормальных случайных величин с неизвестными дисперсиями. Для этого можно использовать статистическую процедуру *Парный двухвыборочный t-тест для* средних.

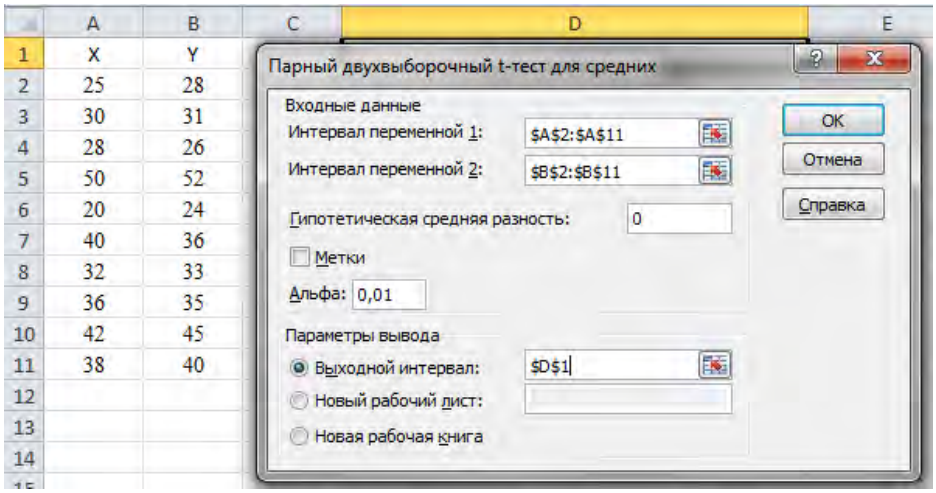


Рис. 4.17 – Исходные данные задачи и диалоговое окно процедуры *Парный двухвыборочный t-тест для средних*

На рис. 4.17 изображено диалоговое окно этой процедуры. Она полностью идентично диалоговому окну процедуры *Двухвыборочный t-тест с одинаковыми дисперсиями.*

Результаты проверки гипотезы приведены на рис. 4.18 в диапазоне ячеек D1:F14.

Анализ результатов решения показывает, что наблюдаемое значение статистики T (–1,132) находится в области принятия гипотезы (–3,25; 3,25). Это означает, что гипотезу о равенстве средних можно принять, т.е. результаты взвешиваний различаются незначимо.

К такому же выводу приводит и сравнение значимости $p = 0{,}287$ с заданным уровнем значимости $\alpha = 0{,}05$: гипотезу H_0 следует принять, так как $p > \alpha$.

K21

	A	B	C	D	E	F
1	X	Y		Парный двухвыборочный t-тест для средних		
2	25	28				
3	30	31			*Переменная 1*	*Переменная 2*
4	28	26		Среднее	34,1	35
5	50	52		Дисперсия	78,767	76,222
6	20	24		Наблюдения	10	10
7	40	36		Корреляция Пирсона	0,959	
8	32	33		Гипотетическая разность средних	0	
9	36	35		df	9	
10	42	45		t-статистика	-1,132	
11	38	40		P(T<=t) одностороннее	0,143	
12				t критическое одностороннее	2,821	
13				P(T<=t) двухстороннее	0,287	
14				t критическое двухстороннее	3,250	

Рис. 4.18 – Проверка гипотезы о равенстве средних результатов взвешиваний

Проверим гипотезу в пакете *Statistica*.

Введём исходные данные.

Для проверки гипотезы будем использовать процедуру Basic *Statistics*/*Tables*, которая находится в меню *Statistics*. Выбираем *t-test, dependent samples* (рис.4.19).

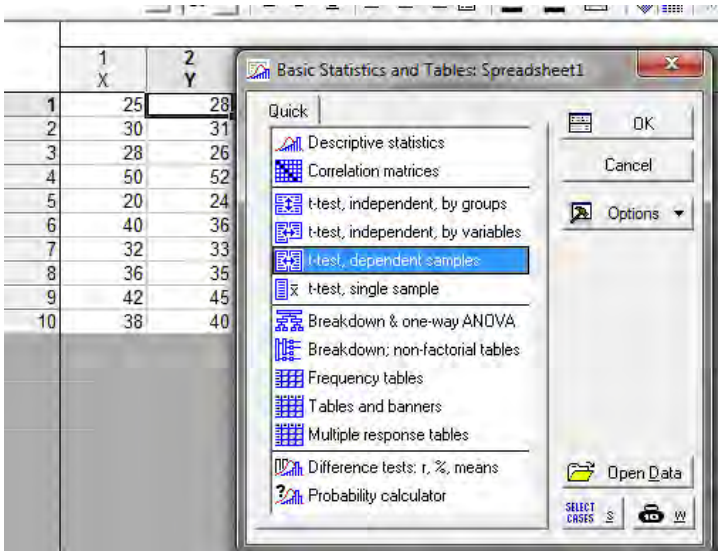


Рис. 4.19 – Выбор *t-test, dependent samples*

Зададим диапазон исходных данных, нажав на кнопку *Variables* и выбрав там *X* и *Y*. После нажатия кнопки OK перейдём на вкладку *Advanced* и зададим уровень значимости α.

Нажав кнопку *Summary*, получаем таблицу (рис. 4.20).

Variable	T-test for Dependent Samples (Spreadsheet1) Marked differences are significant at p < ,01000 Mean	Std.Dv.	N	Diff.	Std.Dv. Diff.	t	df	p
X	34,10000	8,875059						
Y	35,00000	8,730534	10	-0,900000	2,514403	-1,13190	9	0,286933

Рис. 4.20 – Проверка гипотезы о равенстве средних результатов взвешиваний

Из заголовка таблицы следует, что для наличия значимых различий статистическая значимость (*p*-значение) должна быть меньше 0,01 (в нашем случае $p = 0{,}286933$). Значит, гипотеза о равенстве средних результатов взвешиваний принимается – что подтверждает вывод, сделанный в пакете *Excel*.

Контрольный пример 4.5. Ниже приведены результаты измерений производительности 6 агрегатов (по каждому агрегату сделано 5 измерений). Проверить, используя критерии Кохрена, на уровне значимости $\alpha = 0{,}05$ гипотезу о равенстве дисперсий шести наборов данных, характеризующих производительность агрегатов. Задание выполнить в пакетах Excel и MathCad.

Агрегаты					
1	**2**	**3**	**4**	**5**	**6**
14	14,1	14	14,5	12,5	14
14,5	10,1	12,3	14,2	12,3	14
13.7	14.7	12.8	15	11.5	13.5
12.7	13.7	11	14.7	12.9	14.7
14.1	14	13.1	13.5	12.8	13.6

Решение.

1) Работаем в пакете Excel.

На рис. 4.21 приведены результаты проверки гипотезы $H_0 : \sigma_1^2 = K = \sigma_6^2$ о равенстве дисперсий шести наборов данных, характеризующих производительность агрегатов, с помощью критерия Кохрена (в пакете Excel).

В диапазоне $A8 : F8$ находятся выборочные дисперсии шести «наборов» данных, характеризующих производительность данных. Дисперсии вычислены с помощью формулы ДИСП.В(A3:A7), введённой первоначально в ячейку A8 и скопированной затем в ячейки $B8 : F8$.

I5 | fx | =БЕТА.ОБР(1-I4/I1;I2/2;I2*(I1-1)/2)

DA.xlsx *

	A	B	C	D	E	F	G	H	I
1	Агрегаты							l =	6
2	**1**	**2**	**3**	**4**	**5**	**6**		k =	4
3	14	14,1	14	14,5	12,5	14		g =	0,570
4	14,5	10,1	12,3	14,2	12,3	14		α =	0,05
5	13,7	14,7	12,8	15	11,5	13,5		$g(0.05, l, k)$ =	0,480
6	12,7	13,7	11	14,7	12,9	14,7			
7	14,1	14	13,1	13,5	12,8	13,6			
8	**0,46**	**3,372**	**1,223**	**0,327**	**0,31**	**0,223**			

Рис. 4.21. Проверка гипотезы о равенстве нескольких дисперсий с помощью критерия Кохрена

В ячейке I3 находится выборочное значение статистики G, найденное с помощью формулы $=\text{B8}/\text{СУММ}(\text{A8}:\text{F8})$, а в ячейке I5 – критическое значение $g(0{,}05;6;4)=0.4803$ этой статистики, вычисленное с помощью встроенной функции

$$=\text{БЕТА.ОБР}\left(1-\frac{\alpha}{l};\frac{n-1}{2};\frac{l\cdot(n-1)}{2}\right).$$

Полученный результат ($G_{\text{набл}} > g_{\text{kp}}$) свидетельствует о том, что гипотеза о равенстве дисперсий противоречит реальным данным наблюдения и её надо отклонить.

«Виновником» отклонения проверяемой гипотезы, по всей видимости, является агрегат 2. Дисперсия $s_2^2=3.372$ производительности этого агрегата больше суммы дисперсий $s_1^2+s_3^2+s_4^2+s_5^2+s_6^2=2.543$ производительности всех остальных агрегатов.

Решим пример в математическом пакете MathCad (см. рис. 4.22).

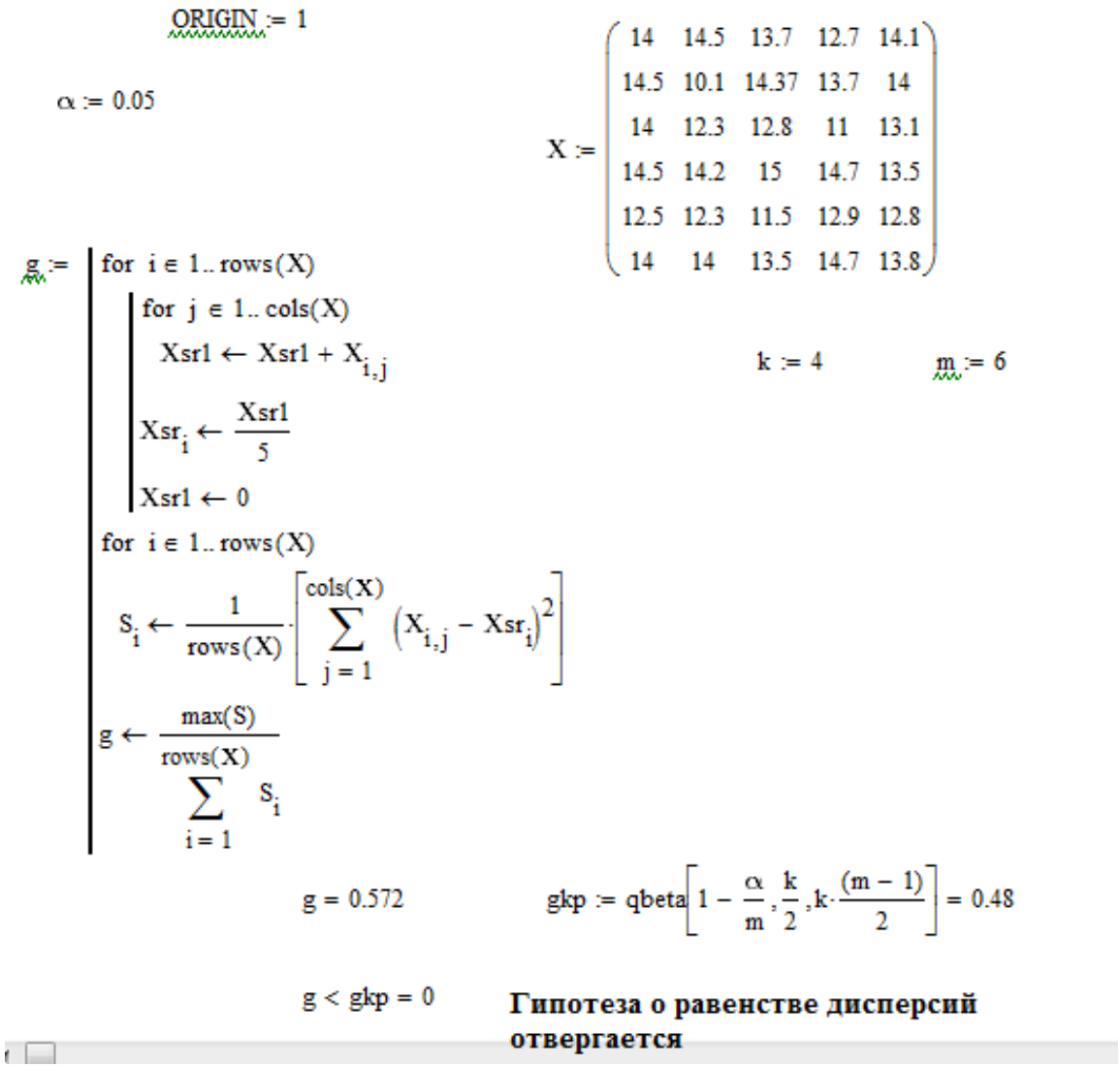


Рис. 4.22. Проверка гипотезы о равенстве нескольких дисперсий с помощью критерия Кохрена в пакете MathCad

Задания для самостоятельной работы

Задание 1. Проектный, контролируемый размер изделий, изготовляемых станком автоматом $a = a_0$ мм. Измерения 20 случайно отобранных изделий дали результаты, приведенные в таблице 4.1.

1. Требуется при уровне значимости $\alpha = 0{,}05$ проверить гипотезу $H_0 : a = a_0$ при конкурирующей гипотезе $H_1 : a \neq a_0$. Задачу решить с применением пакетов Statistica и Excel

2. Партия изделий принимается, если дисперсия контролируемого размера значимо не превышает 0,2. Можно ли принять партию при уровне значимости а) 0,01; б) 0,05? Задачу решить с применением пакетов Excel и Mathcad.

Таблица 4.1

Вариант 1 $a_0 = 30$									
25,8	38,9	24,5	26,9	27,7	24,1	27,9	35,0	29,3	39,1
34,3	28,5	22,2	26,4	26,7	30	30,4	32,3	28,4	35,6
Вариант 2 $a_0 = 24$									
24,5	21,8	23,2	26,4	23,5	26,1	23,1	24,2	25,8	21,7
23,6	28,1	26,2	22,2	25	24,9	23,9	24,5	26,1	24,6
Вариант 3 $a_0 = 35$									
34,6	35,4	34,1	35,3	36,1	33,5	36,2	35,1	35,2	35,5
34,4	33,9	34,8	34,6	34,6	36,1	34,9	35	35,6	34,8
Вариант 4 $a_0 = 30$									
31,3	32,6	29,3	30,6	33,2	28,6	30,5	31,8	29,2	29,6
29,5	28,6	31,7	33,4	36	32,1	28,1	29,0	32,2	34,4
Вариант 5 $a_0 = 34$									
33,5	30,5	32,4	34,6	32,4	36,7	34,8	35,6	40,0	30,3
34,0	32,9	40,0	32,1	33,2	31,7	29,2	31,7	32,5	30,4
Вариант 6 $a_0 = 25$									
24	28	29	21	26	32	19	27	32	25
23	25	26	21	25	21	22	22	21	27
Вариант 7 $a_0 = 33$									
31,5	34,4	32,4	30,1	34,4	31,7	25,0	30,2	33,5	31,7
29,2	26,8	26,4	26,7	33,2	27,2	28,8	30,4	31,6	26
Вариант 8 $a_0 = 26$									
39	34	26	23	28	25	36	25	36	27
20	37	15	31	19	30	24	21	19	17
Вариант 9 $a_0 = 30$									
29,4	30,6	30,6	29,9	29,8	30,7	29,6	29,3	29,9	29,2
29,8	31,4	29,8	30,3	30,6	29,7	28,7	30,2	32,2	30,0
Вариант 10 $a_0 = 24$									
22	24	28	21	22	25	25	30	27	23
23	26	29	16	28	33	28	27	24	24
Вариант 11 $a_0 = 22$									
18,6	21,4	23,0	22,5	21,4	22,4	22,1	19,7	21,5	19,1
22,5	22,9	21,5	21,3	20,6	23,1	23,2	23,1	20,5	22,1
Вариант 12 $a_0 = 40$									
40	38	43	38	41	38	39	36	37	43
36	37	41	38	42	40	44	39	42	39

Задание 2. Для сравнения точности двух станков-автоматов взяты 2 пробы (выборки), объемы которых n_1 и n_2. В результате измерения контролируемого размера отобранных изделий получены результаты, приведенные в таблице 4.2.

1) можно ли считать, что станки обладают одинаковой точностью $\left(H_0 : \sigma_1^2 = \sigma_2^2\right)$, если принять уровень значимости α = 0,05 и в качестве конкурирующей гипотезы $\left(H_1 : \sigma_1^2 \neq \sigma_2^2\right)$. Задание выполнить в пакете *Excel* (с помощью соответствующей процедуры и функции F.TEST).

2) требуется проверить гипотезу $H_0 : a_1 = a_2$ о равенстве средних размеров изделий при конкурирующей $H_1 : a_1 \neq a_2$ не равно. Задание выполнить в пакете *Excel* и *Statistica* (если гипотеза о равенстве дисперсий принимается) и в пакетах *Excel* и *Mathcad*, если отвергается.

Таблица 4.2

Вариант 1														
X	8.2	8.9	9	8.2	8.2	8.2	8.3	7.8	8.5	8.3	8.6	8.6		
Y	8.1	8.4	7.7	8.2	8.1	8.6	7.6	8.2	8.8	8.5	6.9	8.4	8.1	9.2
Вариант 2														
X	1,2	1,4	1,8	2,2	2	2,4	2,1	1,7	1,8	2	2,7	3	3,1	3,3
Y	1,5	2,5	2,3	2,9	2,6	2,8	2,6	2,5	1,9	1,6				
Вариант 3														
X	1,91	1,78	2,78	1,99	1,56	2,28	1,47	2,66	2,2	1,59	2,49			
Y	1,9	2,27	1,82	1,61	2,12	1,94	1,73	20,8	1,96	2,15	2,09			
Вариант 4														
X	72	84	69	74	82	67	75	86	68	61				
Y	55	65	73	66	58	71	77	68	68	59				
Вариант 5														
X	2,42	2,5	1,44	1,94	1,86	2,05	2,21	1,96	2,29	2,31				
Y	2,34	1,66	2,17	1,89	1,76	2,21	2,12	1,88	2,25	2,1				
Вариант 6														
X	3,82	3,64	3,77	3,61	3,79	3,78								
Y	3,95	3,87	3,78	3,86	3,92	3,91	3,89	3,92						
Вариант 7														
X	50.7	50.5	50	50	51.1	50.1	50.1	50.4	49.9	50.9	49.9	50.1		
Y	51.1	50.3	51.2	50.6	50.2	49.8	50.2	49.8	51.1	50.2	50.5	50		
Вариант 8														
X	0,5	0,6	1,4	0,8	1	1,8	0,2	0,4	0,1	0,3	1,1	0,9	0,7	1,3
Y	0,7	0,4	1,4	0,6	0,5	1,3	0,3	1,2	0,2	1				
Вариант 9														
X	12	14	13	16	11	9	13	15	15	18	14			
Y	13	9	11	10	7	6	8	10	11					

Вариант 10														
X	2,5	3,6	2,4	2,8	4	3,9	3,2	2,4	3,1	2,3	4,1	2,9		
Y	2,7	3,1	2,5	4,2	3,9	3,3	2,2	3,6	2,4					
Вариант 11														
X	180	184	185	185	178	184	181	180	181	182	182	178	183	178
Y	184	184	184	185	183	182	185	181	185					
Вариант 12														
X	48	36	28	46	36	24	50	38	26					
Y	39	21	44	31	26	36	24	16	20					

Задание 3. По приведённым ниже выборкам на уровне значимости 0,05 проверить гипотезу о равенстве двух средних нормальных совокупностей с неизвестными дисперсиями при конкурирующей гипотезе $\mathrm{H}_1 : a_1 \neq a_2$.

Предполагается, что случайные величины X и Y распределены нормально и выборки зависимы. Задачу решить с применением пакетов Statistica и Excel.

1	*X*	2.2	2.37	5.67	2.35	2.59	11.7	2.13	2.45	2.55	2.5
	Y	2.1	2.85	2.67	2.34	2.58	2.6	2.16	2.47	2.59	2.4
2	*X*	4.1	3.6	2.2	1.4	3.5	1.6	3.4	1.1	3	2.4
	Y	4.7	5.1	1.8	1.8	3.6	2.5	2.9	1.9	4.9	2.8
3	*X*	7.3	6.2	7.63	6.34	7.71	6.2	4.13	5.45	5.57	5.3
	Y	7.6	6.42	5.62	5.37	7.9	6.6	4.15	6.47	5.59	5.6
4	*X*	15	20	16	22	24	14	18	20		
	Y	15	22	14	25	29	16	20	24		
5	*X*	76	71	57	49	70	69	26	65	59	
	Y	81	85	52	52	70	63	33	83	62	
6	*X*	16	14	14	23	11	12	17	14	18	16
	Y	13	10	11	21	6	9	16	10	16	13
7	*X*	3.5	3.6	7.8	9.6	5.7	8.9	6.3	8.3	4.5	
	Y	1	2.7	8.9	6.5	8.9	6.5	12.5	10.2	1.2	
8	*X*	63	72	85	97	82	101	73	62	58	75
	Y	68	80	95	93	80	106	82	78	65	63
9	*X*	50.2	50.3	50.9	51.1	50.2	50.1	50.9	49.9	50.8	50.9
	Y	50.2	49.9	51.2	50.6	51.1	49.8	50.3	50.9	50.6	50.1
10	*X*	5.03	4.98	5.12	5.08	4.98	5.02	5.01	5.09	5.09	5.07
	Y	5.1	4.8	5.1	5.12	4.99	5.08	5.03	5.02	5.02	5.03
11	*X*	0,1	1,8	2,6	8,6	3,4	3,1	4,4	3	3,7	4,5
	Y	0,8	2,1	2,5	7,2	4,3	3,4	3,3	4,8	3	3,5
12	*X*	0,68	0,18	0,62	0,44	0,73	0,54	1,18	0,56	1,11	0,25
	Y	0,88	0,96	1,07	0,92	0,84	1,12	0,51	1,06	0,93	1,04

Задание 4. С l автоматов, настроенных на обработку одних и тех же деталей, взято по одной текущей выборке объема n. Требуется определить, одинаковая ли точность автоматов, т.е. можно ли принять гипотезу о равенстве дисперсий. Принять $\alpha = 0{,}05$. Задачу решить математическими с применением пакетов *Mathcad* и *Excel.*

Вариант 1								
Номер автомата	**Результаты измерений**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	50	51,2	51	50,7	50	50,4	51	51
2	50,5	50	49,9	51,2	50,4	51,2	50	50,8
3	50,4	50,1	50,9	51,2	51,2	51,1	50	51,1
4	50,4	49,8	50	49,8	51,2	50,7	50	50,8
5	50,5	49,9	50,8	50,2	51,2	50,7	51	50,6

Вариант 2										
Номер автомата	**Результаты измерений**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	58,2	58,4	75	50,7	77,6	46,3	61,2	51	55,2	59,1
2	69,3	41	70,6	51,2	44,9	26,3	62,5	50,8	46	59,7
3	65,9	52,3	66,4	51,2	69,7	33,8	68,6	51,1	57,5	76,6
4	36,5	43,1	50	69,9	88,3	71,1	22	50,8	57,2	37,1

Вариант 3								
Номер автомата	**Результаты измерений**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	50,7	50,7	50,2	50,7	50,1	51,1	50,7	50,6
2	50,6	50,3	49,8	51,1	50,9	49,9	50,6	50,6
3	50,9	50,5	51	50,2	50,7	50,5	51	50,5
4	51,1	49,8	50	50,1	50,2	49,9	50,1	50,3

Вариант 4								
Номер автомата	**Результаты измерений**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	49,8	50,0	50,1	50,1	50,1	50,1	50,0	49,9
2	50,0	50,1	50,1	50,2	49,9	50,2	50,1	49,5
3	49,9	49,9	49,8	49,9	50,1	49,5	50,0	50,1
4	49,9	50,1	50,0	50,2	49,6	50,1	50,2	49,9

Вариант 5										
Номер автомата	**Результаты измерений**									
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
1	44	44	45	44	45	46	46	43	45	47
2	46	44	43	42	45	45	45	45	44	46
3	43	46	44	46	44	45	47	44	47	45
4	44	46	45	45	45	46	44	46	44	46

Вариант 6								
Номер автомата	**Результаты измерений**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	48	53	51	54	52	53	50	54
2	45	55	50	49	53	51	57	47
3	50	48	53	56	52	53	51	57
4	51	54	52	53	55	51	47	56
5	49	53	55	55	52	51	46	53

Вариант 7									
Номер автомата	**Результаты измерений**								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	32	29	28	25	31	30	31	32	31
2	28	31	31	32	30	32	29	30	31
3	29	30	30	33	28	28	33	29	30
4	28	32	31	34	29	32	32	28	28
5	31	30	28	27	31	29	32	31	30

Вариант 8									
Номер автомата	**Результаты измерений**								
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**
1	44	39	39	46	37	37	44	43	40
2	47	39	42	42	41	42	43	41	40
3	40	32	45	39	39	42	43	41	40
4	41	44	39	39	42	42	45	43	43
5	38	43	44	44	37	39	45	40	43

Вариант 9							
Номер автомата	**Результаты измерений**						
	1	**2**	**3**	**4**	**5**	**6**	**7**
1	50,2	50,4	51,2	50,1	51,1	50,9	50,8
2	49,8	50,7	50,1	50,2	50	51	50,8
3	50,2	50,3	51,2	51	49,8	50,6	51
4	49,9	49,8	50,3	49,9	50,7	50,9	50,7

Вариант 10						
Номер авто-мата	**Номер автомата**					
	1	**2**	**3**	**4**	**5**	**6**
1	49,8	50,5	51,1	50	50,2	49,8
2	50,4	49,8	51	51,2	51,1	51,2
3	50,2	50,2	50,5	49,8	49,8	49,9
4	50,9	50	50,1	50,1	50,3	50,7
5	50,4	51,1	51,1	51,2	50,3	51,1

Вариант 11								
Номер автомата	**Результаты измерений**							
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**
1	50,7	50,7	50,2	50,7	50,1	51,1	50,7	50,6
2	50,6	50,3	49,8	51,1	50,9	49,9	50,6	50,6
3	50,9	50,5	51	50,2	50,7	50,5	51	50,5
4	51,1	49,8	50	50,1	50,2	49,9	50,1	50,3

Вариант 12							
Номер автомата	**Результаты измерений**						
	1	**2**	**3**	**4**	**5**	**6**	**7**
1	50	49,8	50,3	50,9	49,9	50,8	50,4
2	50	49,9	50,9	50,3	50,6	51	51
3	51,1	50,3	51,1	50,7	51,1	51,1	51
4	49,8	50,3	50,8	51	49,8	50,1	50,9

Лабораторная работа № 5. Дисперсионный анализ

Необходимые теоретические сведения для выполнения лабораторной работы находятся в теоретическом разделе – тема 3.3.

Контрольный пример 5.1. Результаты наблюдений за расходом сырья при производстве одинаковой продукции по одной и той же технологии на пяти различных заводах равных мощностей, представлены в следующей таблице:

Таблица 5.1

Месяцы	**Расход сырья**				
	Завод 1	**Завод 2**	**Завод 3**	**Завод 4**	**Завод 5**
1	114	112	132	124	124
2	124	119	124	114	116
3	110	124	129	119	119
4	116	116	129	124	119
5	119	116	129	116	132
6	119	124	124	116	129
7	129	112	114	129	116
8	124	119	119	124	119
9	110	119	124	114	
10	124	112		116	
11	119			129	
12	124				

Известно, что расход сырья является нормально распределённой случайной величиной и дисперсии наблюдений по каждому заводу равны.

При уровне значимости $\alpha = 0{,}05$ требуется выяснить, зависит ли расход сырья от того, на каком заводе произведена продукция. Задания выполнить в пакетах Excel и Statistica.

Решение.

Введём исходные данные на лист MS Excel, как показано на рис. 5.1.

Для каждого завода рассчитаем групповую среднюю (см. рис. 5.1).

	A	B	C	D	E
1	**Завод 1**	**Завод 2**	**Завод 3**	**Завод 4**	**Завод 5**
2	114	112	132	124	124
3	124	119	124	114	116
4	110	124	129	119	119
5	116	116	129	124	119
6	119	116	129	116	132
7	119	124	124	116	129
8	129	112	114	129	116
9	124	119	119	124	119
10	110	119	124	114	
11	124	112		116	
12	119			129	
13	124				
14	**Групповые средние**				
15	**119,33**	**117,30**	**124,89**	**120,45**	**121,75**

Рис. 5.1 – Исходные данные к примеру

В ячейку A15 введём формулу = СРЗНАЧ(A2:A13), которую затем скопируем методом автозаполнения вправо по строке. При этом Excel игнорирует пустые ячейки при расчёте среднего значения (как и при использовании других статистических функций). Общую выборочную среднюю рассчитаем с помощью функции СРЗНАЧ(A2:E13) (рис. 5.2).

G	H	I	J	K	L	M
Общая выборочная средняя						
		120,56				
Групповые суммы квадратов отклонений						
394,67	**186,10**	**256,89**	**328,73**	**251,50**		
Общая сумма квадратов отклонений Qобщ					**1722,32**	
Общее число наблюдений			**Число уровней фактора**			
	50			**5**		

Результаты расчётов по дисперсионному анализу						
	Сумма квадратов отклонений	**Число степеней свободы**	**Дисперсия**	**Fнабл**	**p-значение**	**Fкрит**
межгрупповая	304,44	4	76,11	2,42	0,063	2,58
внутригрупповая	1417,88	45	31,51			
общая	1722,32					
Выборочный коэффициент детерминации				0,18		

Рис. 5.2 – Вид листа MS Excel с расчётами для дисперсионного анализа

Аналогично рассчитаем суммы квадратов отклонений от среднего для каждого завода (рис. 5.2).

Введём формулу =КВАДРОТКЛ(A2:A13) в ячейку G5, которую затем скопируем вправо по строке. Общую сумму квадратов отклонений $Q_{\text{общ}}$ рассчитаем с помощью функции КВАДРОТКЛ(A2:E13).

Общее число наблюдений рассчитаем с помощью функции СЧЕТ(A2:A13), которая подсчитывает число заполненных числами ячеек в заданном диапазоне, а пустые ячейки игнорирует (рис. 5.2).

Результаты расчётов по дисперсионному анализу с помощью функций MS Excel приведены на рис. 5.2.

Общая сумма квадратов отклонений уже была рассчитана, поэтому в ячейке H14 поставим ссылку на ячейку L6.

Внутригрупповую (остаточную) сумму квадратов отклонений $Q_{\text{ост}}$ рассчитаем в ячейке H13, сложив соответствующие значения для всех заводов (=СУММ(G5:K5)).

Межгрупповую (факторную) сумму квадратов отклонений найдём как разность значений в ячейках H14 и H13.

Введём в ячейки I12 и I13 число степеней свободы: для межгрупповой дисперсии это $m-1=5-1=4$, для внутригрупповой дисперсии $n-m=45$.

Рассчитаем межгрупповую и внутригрупповую дисперсии в ячейках J12 и J13, разделив соответствующие суммы квадратов отклонений на число степеней свободы.

Рассчитаем теперь наблюдаемое значение критерия Фишера, разделив межгрупповую (факторную) дисперсию на внутригрупповую (остаточную) дисперсию (ячейка K12). Таким образом, $F_{\text{набл}} \approx 2{,}42$.

Для расчёта критической точки распределения Фишера в ячейке M12 используем функцию Excel F.ОБР.ПХ(0,05; 4; 45) (рис. 5.3).

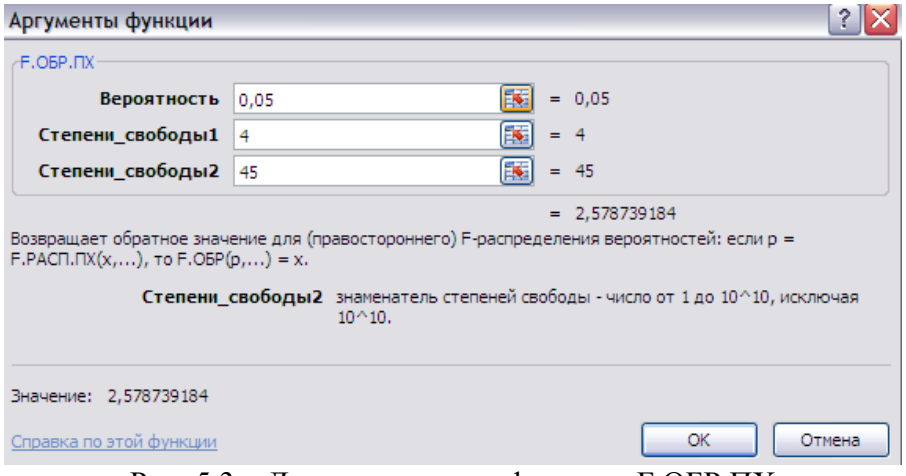


Рис. 5.3 – Диалоговое окно функции F.ОБР.ПХ

Получим $F_{\text{крит}} = 2{,}58$. Так как $F_{\text{набл}} < F_{\text{крит}}$, можно считать несущественным влияние фактора и принять гипотезу о равенстве математических ожиданий генеральных совокупностей, соответствующих каждому заводу. Таким образом, доказано, что расход сырья на производство исследуемого вида продукции не зависит от завода.

Проверку гипотезы можно реализовать и с помощью p – значения. В MS Excel этот показатель рассчитывается с помощью функции F.РАСП.ПХ $(F_{\text{набл}}; k_1; k_2)$. Если полученное число $p > \alpha$, то нулевую гипотезу о равенстве математических ожиданий групп нужно принять (нет влияния фактора). Если же $p < \alpha$, то нулевая гипотеза отклоняется, т.е. признаётся влияние фактора. В нашем случае p – значение рассчитано в ячейке L12. Полученный результат означает, что для всех уровней значимости, меньших либо равных 0,063, гипотеза о равенстве математических ожиданий может быть принята. Поскольку $\alpha = 0{,}05 < 0{,}063$, влияние фактора признаётся несущественным.

Выборочный коэффициент детерминации

$$R^2 = \frac{Q_{\text{факт}}}{Q_{\text{ост}}} = \frac{304{,}44}{1722{,}32} \approx 0{,}18$$

рассчитан в ячейке K12. Он означает, что только 18% общей выборочной вариации расхода сырья связано с выбором завода.

Аналогичные результаты получим с помощью инструмента из *Пакета анализа*.

Зададим команду *Данные/Анализ данных* и выберем *Однофакторный дисперсионный анализ*. Заполним диалоговое окно, как показано на рис. 5.4.

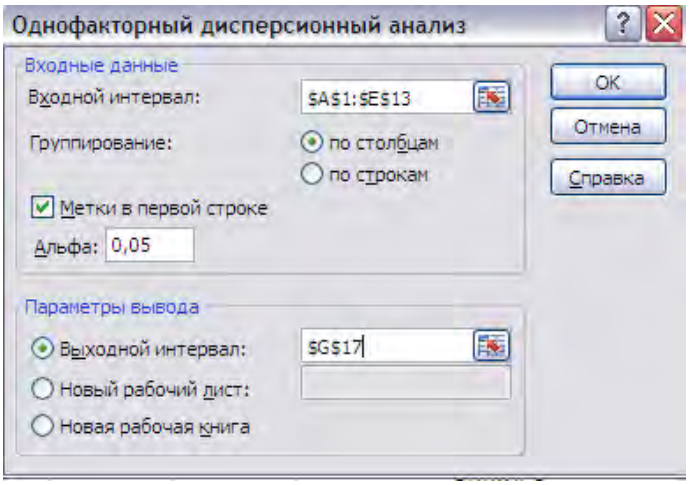


Рис. 5.4. Диалоговое окно *Однофакторный дисперсионный анализ*

Флажок *Метки в первой строке* поставлен потому, что входной интервал A1:E13 включает заголовки столбцов, и они будут использованы для

формирования результата. Результат работы этого инструмента анализа представлен на рис. 5.5.

Однофакторный дисперсионный анализ

ИТОГИ

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
Завод 1	12	1432	119,3333333	35,87878788
Завод 2	10	1173	117,3	20,67777778
Завод 3	9	1124	124,8888889	32,11111111
Завод 4	11	1325	120,4545455	32,87272727
Завод 5	8	974	121,75	35,92857143

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	304,4371717	4	76,10929293	2,415515664	0,062563703	2,578739184
Внутри групп	1417,882828	45	31,5085073			
Итого	1722,32	49				

Рис. 5.5 – Результат работы инструмента *Однофакторный дисперсионный анализ*

В первой таблице результатов анализа показаны выборочные характеристики для каждого уровня фактора: количество наблюдений (счёт), сумма значений, среднее и дисперсия.

Во второй таблице результатов показаны расчёты для дисперсионного анализа, аналогичные тем, что были ранее рассчитаны с помощью стандартных функций MS Excel.

Решение в пакете Statistica. Введём исходные данные из таблицы в созданную таблицу в формате *Statistica*, как показано на рисунке 5.6.

	1 Var1	2 Var2						
1	1	114	18	2	124			
2	1	124	19	2	112			
3	1	110	20	2	119	36	4	116
4	1	116	21	2	119	37	4	116
5	1	119	22	2	112	38	4	129
6	1	119	23	3	132	39	4	124
7	1	129	24	3	124	40	4	114
8	1	124	25	3	129	41	4	116
9	1	110	26	3	129	42	4	129
10	1	124	27	3	129	43	5	124
11	1	119	28	3	124	44	5	116
12	1	124	29	3	114	45	5	119
13	2	112	30	3	119	46	5	119
14	2	119	31	3	124	47	5	132
15	2	124	32	4	124	48	5	129
16	2	116	33	4	114	49	5	116
17	2	116	34	4	119	50	5	119

Рис. 5.6. – Исходные данные.
Var1 – факторы; *Var2* – зависимая переменная.

Проведём анализ в модуле *ANOVA* (Дисперсионный анализ).

Из переключателей модулей *Statistica* откроем модуль *ANOVA*. Высветим название модуля и далее щёлкнем мышью по названию модуля: *ANOVA* (рис 5.7).

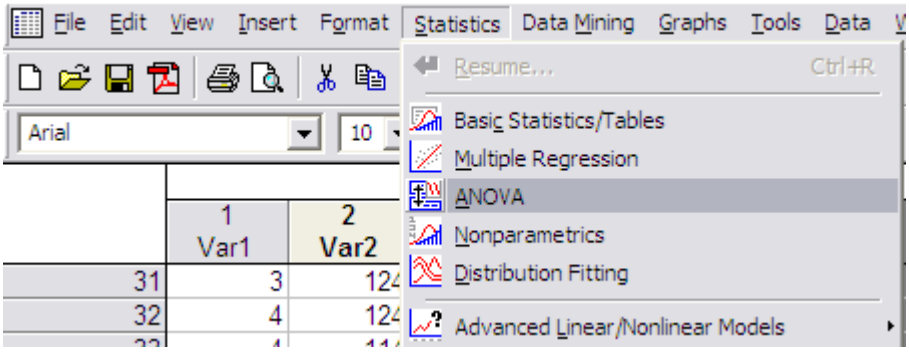


Рис. 5.7 – Основное меню

На экране появится стартовая панель модуля. Выполним установки, как показано на рис. 5.8.

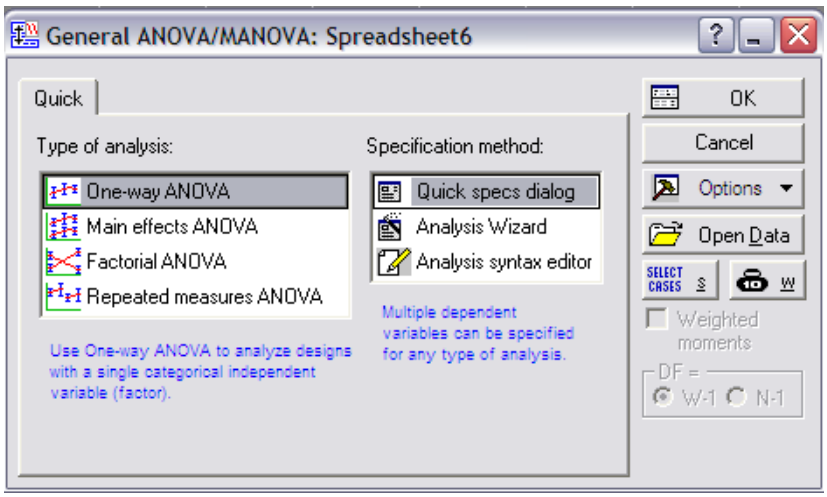


Рис. 5.8 – Стартовая панель модуля

После нажатия кнопки OK в появившемся окне (рис. 5.9) выберем переменные для анализа.

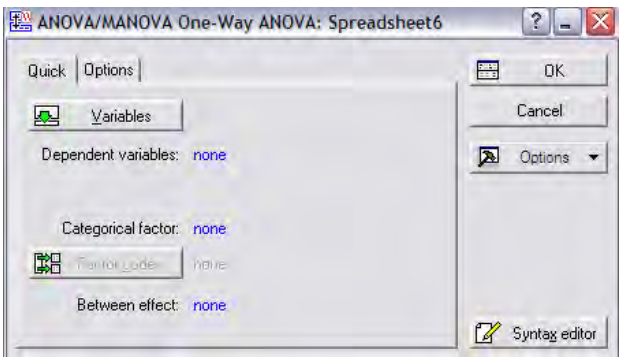


Рис. 5.9 – Выбор переменных

Выбор переменных осуществляется с помощью кнопки *Variables* (Переменные). После того, как кнопка будет нажата, диалоговое окно *Select dependent variables and a categorical predictor (factor)* (Выбрать списки зависимых переменных и факторов) появится на экране (рис. 5.10).

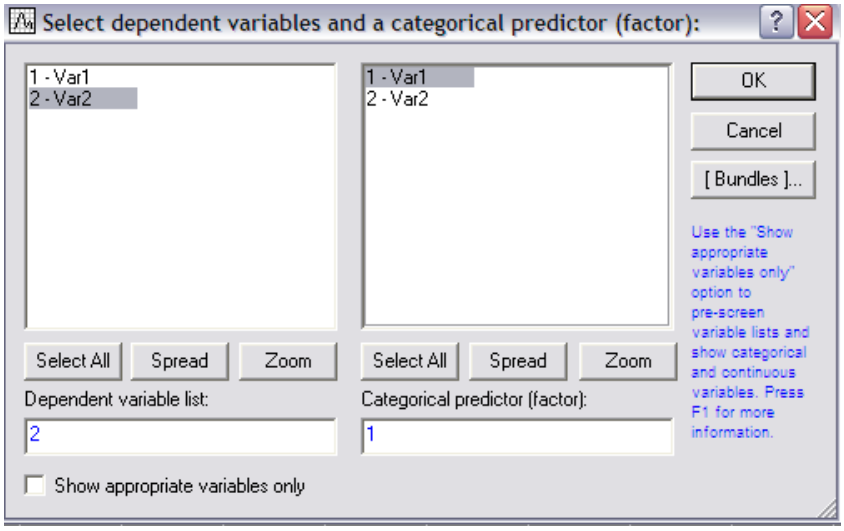


Рис. 5.10 – Окно выбора переменных для анализа

В левой части окна выберем зависимую переменную, а в правой – фактор. После нажатия кнопки OK в появившемся окне выберем *All*, нажав кнопку *Factor codes* (рис. 5.11).

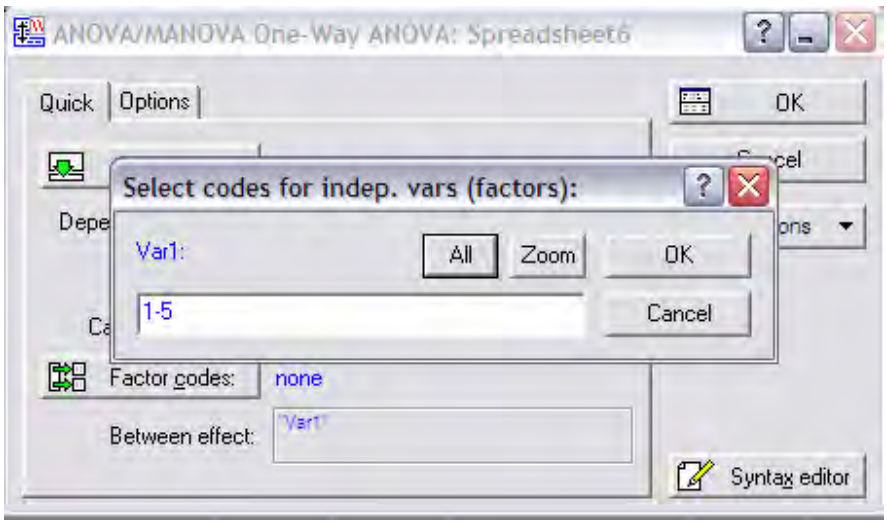


Рис. 5.11 – Окно выбора факторов.

Нажмём кнопку OK в правом углу стартовой панели.

Замечание 5.1. В программе *Statistica* можно проверить выполнение основных предположений, оправдывающих применение дисперсионного анализа. Наиболее важными из них являются два: 1) нормальность распределений по уровням фактора и 2) однородность (или гомогенность) дисперсий.

В левом нижнем углу диалогового окна *Anova Result 1* нажмём клавишу *More results* (Больше), перейдя к развёрнутому представлению результатов (рис. 5.12):

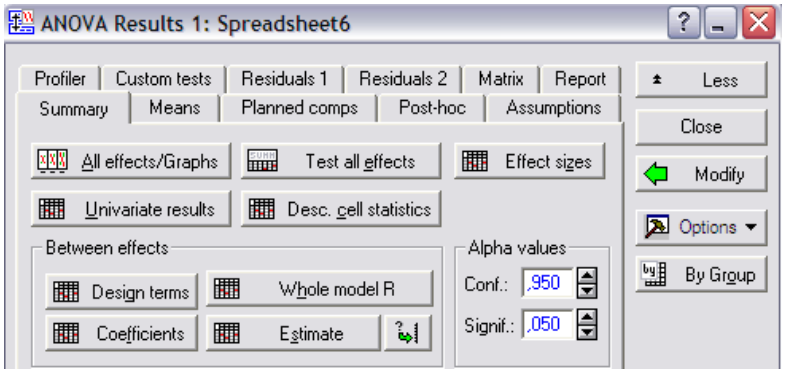


Рис. 5.12. Развёрнутая панель результатов дисперсионного анализа.

Выберем вкладку *Assumptions* (Допущения) и нажмём на кнопку *Cochran C, Hartley, Bartlett* (тест Кохрена, Хартли, Бартлетта). Появится следующая таблица (рис. 5.13):

Tests of Homogeneity of Variances (Spreadsheet6)
Effect: "Var1"

	Hartley F-max	Cochran C	Bartlett Chi-Sqr.	df	p
Var2	1,737545	0,228163	0,829796	4	0,934409

Рис. 5.13 – Проверка дисперсий на однородность

Как видно из таблицы, проверка дисперсий на однородность осуществляется одновременно по трём тестам.

Так как *p*-значение больше 0,05, то принимается нулевая гипотеза, и дисперсии подвыборок, сформированных по уровням фактора – однородны. Если дисперсии неоднородны, то дисперсионный анализ проводить не стоит.

Замечание 5.2. Критерий Хартли, так же как и критерий Кохрена, используется в случае выборок равного объёма, в то время как критерий Бартлетта может применяться и для выборок, чей объём различен (как в нашей задаче).

Вернёмся в диалоговое окно *Anova Result 1* и выберем на вкладке *Summary – Univariate Results* (Результат дисперсионного анализа). В окне результатов рисунка 5.14 представлены результаты дисперсионного анализа:

– между группами – *Var1*;

– внутри групп – *Error*.

Effect	Univariate Results for Each DV (Spreadsheet6) Sigma-restricted parameterization Effective hypothesis decomposition Degr. of Freedom	Var2 SS	Var2 MS	Var2 F	Var2 p
Intercept	1	714183,4	714183,4	22666,37	0,000000
"Var1"	4	304,4	76,1	2,42	0,062564
Error	45	1417,9	31,5		
Total	49	1722,3			

Рис. 5.14. Результаты дисперсионного анализа.

Различие между средними статистически незначимо (на уровне 0,062564, то есть больше, чем критическое значение 0,05). Поскольку различие между средними значениями незначимо, нулевая гипотеза принимается (результат в строке: между группами – *Var1* подсвечивается чёрным цветом).

Контрольный пример 5.2. Дана информация о среднем потреблении топлива на 100 километров в литрах в зависимости от объема двигателя и вида топлива:

	Бензин со свинцом	Бензин без свинца	Дизельное топливо	Среднее $\bar{x}_i$
1001-1500 см^3	9,3	8,9	6,5	**8,23**
1501-2000 см^3	9,4	9,1	7,1	**8,53**
Более 2000 см^3	12,6	9,8	8	**10,13**
Среднее $\bar{x}_j$	**10,43**	**9,27**	**7,2**	

Требуется проверить, зависит ли потребление топлива от объема двигателя и вида топлива.

Решение.

1. В пакете *Excel*:

Введем исходные данные (рис. 5.15):

	A	B	C	D
1		Бензин со свинцом	Бензин без свинца	Дизельное топливо
2	1001-1500 см^3	9,3	8,9	6,5
3	1501-2000 см^3	9,4	9,1	7,1
4	Более 2000 см^3	12,6	9,8	8

Рис. 5.15 – Исходные данные задачи

Выведем на экран диалоговое окно *Двухфакторный дисперсионный анализ без повторений* (рис. 5.16).

В поле *Входной интервал* введем ссылку на диапазон ячеек, содержащий исходные данные. Установим флажок *Метки*. Оставим без изменений предлагаемый процедурой уровень значимости $\alpha = 0{,}05$. Щелчком на переключателе Выходной интервал активизируем поле ввода, находящееся справа от

этого переключателя, и введем в него ссылку на левую верхнюю ячейку таблицы результатов решения.

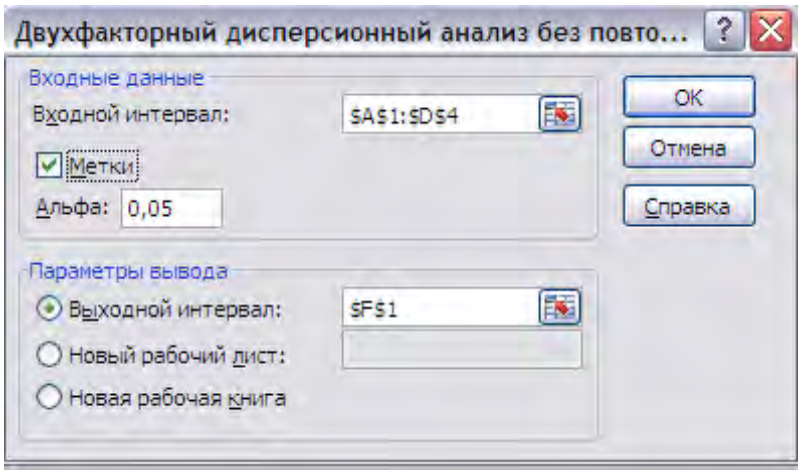


Рис. 5.16 – Диалоговое окно процедуры
Двухфакторный дисперсионный анализ без повторений.

Щелкнем на кнопке ОК. Справа от таблицы исходных данных появятся 2 таблицы результатов рассматриваемой процедуры (рис. 5.17):

E	F	G	H	I	J	K	L
	ИТОГИ	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
	1001-1500 см^3	3	24,7	8,233	2,293		
	1501-2000 см^3	3	25,6	8,533	1,563		
	Более 2000 см^3	3	30,4	10,133	5,373		
	Бензин со свинцом	3	31,3	10,433	3,523		
	Бензин без свинца	3	27,8	9,267	0,223		
	Дизельное топливо	3	21,6	7,2	0,57		
	Дисперсионный анализ						
	Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
	Строки	6,26	2	3,13	5,2753	0,0756	6,9443
	Столбцы	16,08667	2	8,0433	13,5562	0,0165	6,9443
	Погрешность	2,373333	4	0,5933			
	Итого	24,72	8				

Рис. 5.17 – Результаты решения контрольного примера 5.2.

Фактор *A* (объем двигателя) сгруппирован в строках. Так как фактическое отношение Фишера 5,275 меньше критического 6,944, с вероятностью 95% принимаем, что потребление топлива не зависит от объема двигателя.

Фактор *B* (вид топлива) сгруппирован в столбцах. Фактическое отношение Фишера 13,556 больше критического 6,944, поэтому с вероятностью 95% принимаем, что потребление топлива зависит от его вида.

2. В пакете Statistica работаем в модуле ANOVA:

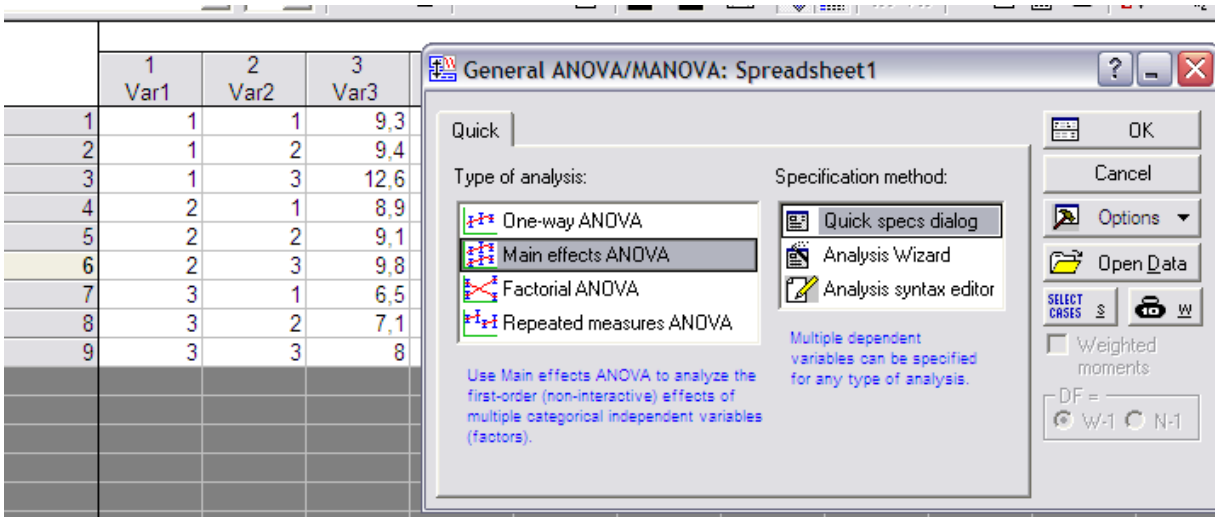


Рис. 5.18 – Исходные данные и вызов модуля.

В нем выбираем пункт *Quick Specs Dialog* в колонке *Specification Method* и *Main effects ANOVA* в колонке *Type of analysis*.

Нажимаем *OK*. Открывается окно *Variables*. Нажмем кнопку *OK* и определим зависимую (VAR3) и независимые (VAR1 – VAR2) переменные:

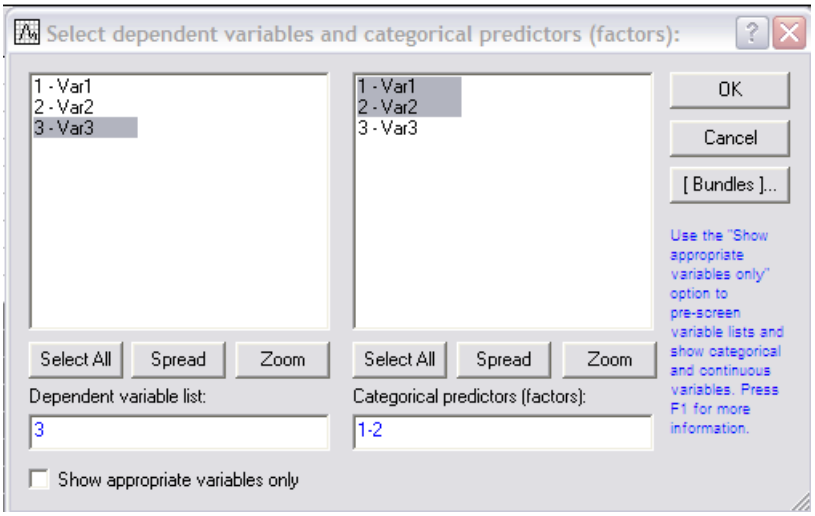


Рис. 5.19 – Выбор переменных.

После нажатия кнопки OK в появившемся окне выберем *All*, нажав кнопку *Factor codes*. Далее необходимо нажать **ОК**. Появится панель *ANOVA Results*. Нажимаем кнопку *All effect*:

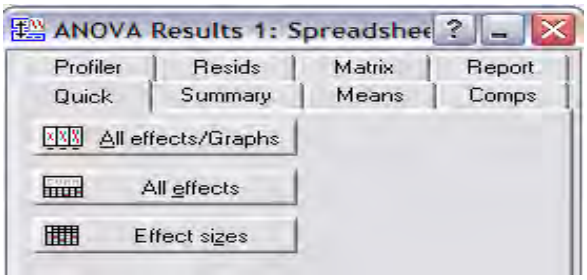


Рис. 5.20 – Панель *ANOVA Results*

В открывшемся окне мы получаем результат решения нашей задачи

Univariate Tests of Significance for Var3 (Spreadsheet1
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	SS	Degr. of Freedom	MS	F	p
Intercept	723,6100	1	723,6100	1219,567	0,000004
"Var1"	16,0867	2	8,0433	13,556	0,016529
"Var2"	6,2600	2	3,1300	5,275	0,075572
Error	2,3733	4	0,5933		

Рис. 5.21 – Результаты дисперсионного анализа

В рассмотренном примере *F*-критерий показывает, что различие между средними статистически значимо (значимо на уровне 0,016529, то есть меньше, чем критическое значение 0,05). Поскольку различие между средними значениями значимо, нулевая гипотеза отвергается и принимается альтернативная гипотеза о существовании различия между средними (результат в строке: между группами – Var1 подсвечивается красным цветом).

Подтверждается вывод, сделанный при выполнении работы в *Excel:* принимаем, что потребление топлива не зависит от объема двигателя; но зависит от его вида

Задания для самостоятельной работы

Задание 1. Для заданного уровня значимости $\alpha = 0{,}05$ установить влияние типа используемой рекламы на объём продаж товара. Определить степень влияния типа используемой рекламы на объём продаж товара. Задание выполнить и в пакетах *Statistica* и *Excel*.

Вариант 1

Тип рекламы	**Годы**					
	1	**2**	**3**	**4**	**5**	**6**
A	215	224	222	221	206	207
B	215	222	207	214	224	217
C	222	237	221	233	229	232
D	242	242	238	250	239	

Вариант 2

Тип рекламы	Годы					
	1	2	3	4	5	6
A	207.33	221.83	241.34	223.13	221.83	221.6
B	239.27	216.28	220.53	207.15	209.76	235.13
C	217.8	222.71	225.32	193.85	216.37	
D	256.51	243.92	234.56	244.94		

Вариант 3

Тип рекламы	Годы					
	1	2	3	4	5	6
A	224.85	238.64	222.86	221.81	231.95	215.19
B	215.31	205.01	223.35	189.37	224.08	224.87
C	232.26	240.63	226.25	215.57	233.51	
D	247.09	238.2	238.91	235.53		

Вариант 4

Тип рекламы	Годы					
	1	2	3	4	5	6
A	232.35	230.84	225.14	216.2	240.17	222.34
B	225.61	222.26	220.4	235.2	218.61	215.31
C	248.46	229.51	214.23	215.34	235.54	
D	224.86	231.9	238.82	243.59		

Вариант 5

Тип рекламы	Годы					
	1	2	3	4	5	6
A	215.19	224.97	235.58	211.15	222.19	217.20
B	213.61	236.47	238.55	212.78	224.66	229
C	212.35	214.91	221.44	227.28	230.34	
D	215.42	208.4	227.59	229.52		

Вариант 6

Тип рекламы	Годы					
	1	2	3	4	5	6
A	221.82	233.26	225.66	225.59	228.08	213.07
B	205.38	226.84	218.96	216.52	223.42	234.16
C	248.91	221.06	220.78	251.27	219.37	
D	244.76	239.35	233.18	243.94		

Вариант 7

Тип рекламы	Годы					
	1	2	3	4	5	6
A	228.08	213.07	201.07	232.94	235.75	229.09
B	223.42	234.16	205.96	234.04	231.84	240.47
C	219.37	217.77	218.48	234.35	232.74	
D	227	225.29	243.25	240.81		

Вариант 8

Тип рекламы	Годы					
	1	2	3	4	5	6
A	235.75	229.09	215.3	220.5	248.46	218.27
B	213.84	240.47	225.63	214.86	229.73	243.61
C	232.74	225.62	252.79	244.56	234.89	
D	241.19	225.51	248.01	236.42		

Вариант 9

Тип рекламы	Годы					
	1	2	3	4	5	6
A	220.15	235.6	234.69	213.86	242.61	229.8
B	201.31	226.23	208.01	212.02	219.59	221.27
C	218.54	219..72	236.49	230.72	214.55	
D	242.49	230.79	252.28	258.23		

Вариант 10

Тип рекламы	Годы					
	1	2	3	4	5	6
A	242.61	229.8	218.36	223.89	236.38	235.37
B	203.74	242.1	220.57	229.62	224.19	210.96
C	220.63	210.11	223.19	234.36	234.87	
D	234.83	243.85	244.86	244.71		

Вариант 11

Тип рекламы	Годы					
	1	2	3	4	5	6
A	224.19	210.96	233.7	220.25	215.96	205.11
B	212.19	229.79	235.49	220.67	212.68	226.58
C	234.87	242.22	237.09	223.85	224.41	
D	243.65	233.64	244.29	249.91		

Вариант 12

Тип рекламы	Годы					
	1	2	3	4	5	6
A	232.63	220.86	216.7	216.35	204.55	220.24
B	205.38	207.78	228.74	232.09	219.36	228.82
C	237.68	227.69	224.84	239.95	218.23	
D	243.08	241.23	243.86	219.11		

Задание 2. Задание выполнить в пакетах *Statistica* и *Excel*.

Варианты 1–5. В двухфакторном комплексе приводится сменная выработка рабочего в зависимости от типа станка (*A*) и стажа его работы (*B*). При $\alpha=0.01$ проверить влияние факторов *A* и *B* на сменную выработку рабочего (таблица 5.6).

Таблица 5.6

	A_1	A_2	A_3
Вариант 1			
B_1	195	198	202
B_2	196	201	203
B_3	198	202	204
Вариант 2			
B_1	208	203	202
B_2	192	195	193
B_3	198	201	203
Вариант 3			
B_1	195	198	202
B_2	197	208	206
B_3	192	190	195
Вариант 4			
B_1	189	188	179
B_2	186	190	193
B_3	198	203	201
Вариант 5			
B_1	199	191	189
B_2	204	201	203
B_3	208	202	204

Варианты 6 – 10. При исследовании зависимости товарооборота центральной районной аптеки от товарооборота (A) и штатной численности прикрепленной аптечной сети (В) получен двухфакторный комплекс. При α=0.05 проверить существенность влияния факторов A и B на товарооборот (таблица 5.7).

Таблица 5.7

	A_1	A_2	A_3
Вариант 6			
B_1	157	163	161
B_2	160	165	158
B_3	158	163	158
Вариант 7			
B_1	152	151	154
B_2	144	145	133
B_3	131	135	138

Продолжение таблицы 5.7

Вариант 8			
B_1	122	128	126
B_2	128	118	116
B_3	162	160	165
Вариант 9			
B_1	159	158	160
B_2	166	160	163
B_3	158	153	156
Вариант 10			
B_1	159	161	159
B_2	144	146	143
B_3	128	122	128

Варианты 11–12. Фактор A имеет 4 уровня, фактор B – 5 уровней. Сделано по одному измерению случайной величины X на каждой комбинации уровней факторов. Полученные результаты представлены в следующей таблице:

Таблица 5.9

	A_1	A_2	A_3	A_4
Вариант 11				
B_1	19	25	17	21
B_2	22	19	19	18
B_3	26	23	22	25
B_4	18	26	20	23
B_5	21	22	21	24
Вариант 12				
B_1	38	44	32	31
B_2	32	41	33	36
B_3	46	45	40	38
B_4	44	42	37	36
B_5	35	33	33	34

Лабораторная работа № 6. Корреляционный анализ

Необходимые теоретические сведения для выполнения лабораторной работы находятся в теоретическом разделе – тема 4.1.

Контрольный пример 6.1. Для исследования зависимости случайных величин X и Y получены статистические данные, представленные в таблице (табл. 6.1).

Таблица 6.1.

X	0	1	2	4	6	8	9	10
Y	6	7,2	9,4	11	15,2	16,6	19,4	21,2

Требуется:

1) Построить корреляционное поле, сделать вывод.
2) Найти корреляционный момент и выборочный коэффициент корреляции.
3) При уровне значимости $\alpha = 0{,}05$ проверить нулевую гипотезу о равенстве генерального коэффициента корреляции нулю при конкурирующей гипотезе $H_1 : r_{XY} \neq 0$.

Задание выполнить в пакетах *Statistica* и *Excel*.

Решение.

1) Выполнение в пакете *Statistica*.

Сначала нужно заполнить таблицу данных (аналог корреляционной таблицы), на основе которой будет проводиться анализ. Для этого после открытия приложения *Statistica* в меню *File* выбираем пункт *New* для создания нового документа. В результате появляется окно, где в графе *Number of variables* указываем количество переменных (в нашем случае 2), а в графе *Number of cases* – количество значений (в примере – 8).

В результате открывается окно, в которое можно вносить различные значения для переменных (рис. 6.1).

	1 x	2 y
1	0	6
2	1	7,2
3	2	9,4
4	4	11
5	6	15,2
6	8	16,6
7	9	19,4
8	10	21,2

Рис. 6.1 – Исходные данные

Построим корреляционное поле, выбрав в меню *Graphs* (Графики) пункт *Scatterplots* (Диаграммы рассеяния). Далее выбираем вкладку *Advanced:*

Выберем *Graph type/Regular*; *Fit/Off* и нажмём кнопку OK, после чего на экране появляется график корреляционного поля (рис. 6.2):

Далее открываем меню *Statistics* и выбираем пункт *Basic Statistics and Tables* (Основные статистики). После этого в появившемся окне открываем пункт *Correlation matrices* (Корреляционная матрица).

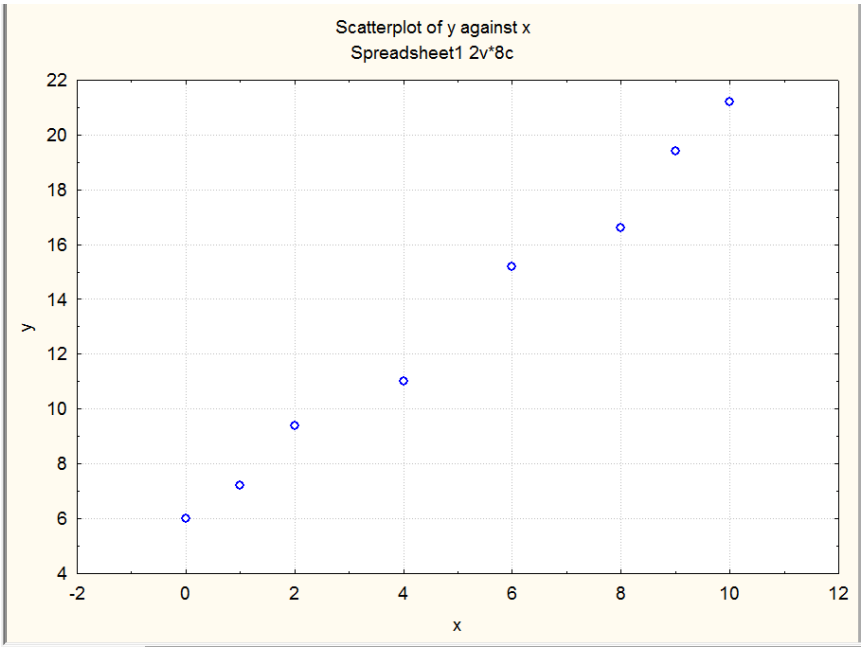


Рис.6.2 – Корреляционное поле

Открывшийся пункт позволяет рассчитать коэффициент корреляции Пирсона. Для этого нажимаем кнопку *Two lists* и в появившемся окне последовательно выбираем переменные, для которых нужно определить зависимость (рис.6.3).

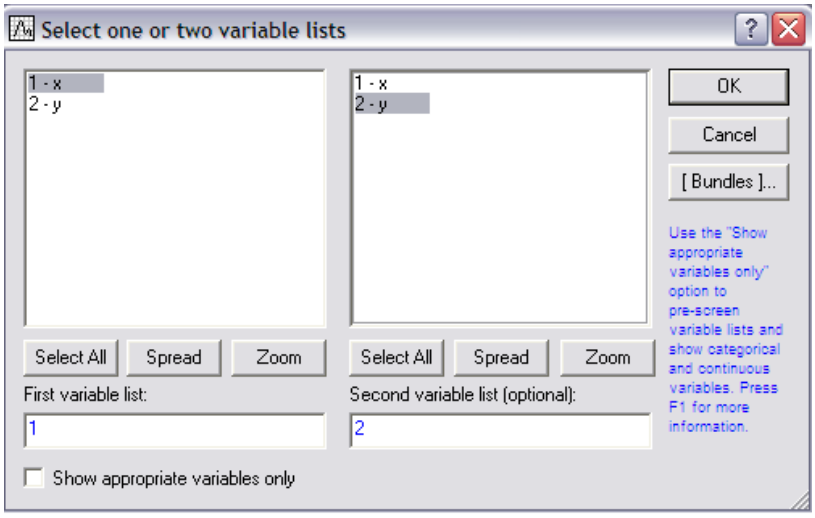


Рис. 6.3 – Выбор переменных для корреляционного анализа

После нажатия кнопки OK вернёмся в предыдущее окно, где производим дальнейшие настройки. Во вкладке *Options*(опции), в зависимости от того, какой вид отчета нужно вывести, выбираем один из пунктов: *Display simple matrix* (вывод только посчитанного коэффициента Пирсона), *Display r, p-levels and N's* (вывод коэффициента, уровня значимости α и числа значений переменной) или *Display detailed table of results* (вывод детального отчета). В этой же вкладке можно изменять уровень значимости для расчетов, изменяя

значения пункта *p-level for highlighting*. Выберем *Display detailed table of results*.

Когда все настройки закончены, нажимаем кнопку *Summary* для вывода результатов. В результате на экран выводится результат анализа (рис. 6.4):

Correlations (Spreadsheet1)
Marked correlations are significant at p < ,05000
(Casewise deletion of missing data)

Var. X & Var. Y	Mean	Std.Dv.	r(X,Y)	r?	t	p
x	5,00000	3,817254				
y	13,25000	5,670727	0,993887	0,987812	22,05188	0,000001

Рис. 6.4 – Расчет коэффициента выборочной корреляции

Так как $p < \alpha$, гипотеза о равенстве нулю генерального коэффициента корреляции отвергается.

2) Выполнение в пакете *Excel*.

Введём исходные данные (рис. 6.5). Зададим команду *Данные – Анализ данных* и выберем инструмент *Корреляция.* Заполним диалоговое окно «Корреляция» как показано на рис. 6.5.

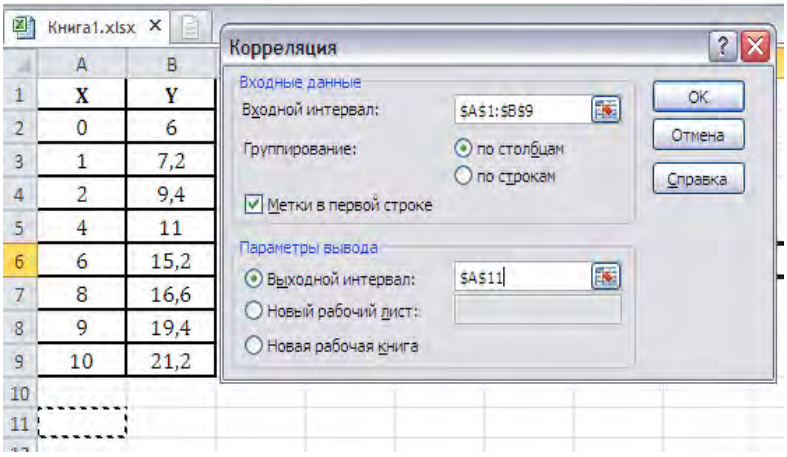


Рис. 6.5 – Исходные данные и пример заполнения диалогового окна «Корреляция»

Входной интервал охватывает все фактические данные, причём каждой случайной отведён отдельный столбец, на что указывает переключатель *Группирование*. Первая строка, содержащая заголовки столбцов, также вошла в диапазон входного интервала, поэтому был установлен флажок *Метки в первой строке*. Переключатель *Параметры вывода* установлен в положение *Выходной интервал*, чтобы корреляционная матрица расположилась на текущем листе.

Коэффициент линейной корреляции равен $0{,}993887 > 0{,}7$ (рис. 6.7), что свидетельствует о весьма высокой степени прямой линейной связи (по шкале

Чаддока). Проверим его на значимость. Найдём наблюдаемое значение T – статистики.

Введём в ячейку E11 формулу MS Excel:

=ABS(B13)/КОРЕНЬ(1–B13^2)*КОРЕНЬ(8–2).

Для расчёта критического значения T – статистики при уровне значимости $\alpha = 0{,}05$ и числе степеней свободы $n-2=6$ введём в ячейку E13 формулу

=СТЬЮДЕНТ.ОБР.2Х (0, 05; 6).

Результаты показаны на рис. 6.6.

	A	B	C	D	E	F	G
10					**наблюдаемое значение t-критерия**		
11		*X*	*Y*		**22,052**		
12	X	1			**критическое значение t-критерия**		
13	Y	0,9939	1		**2,4469**		
14							

Рис. 6.6 – Корреляционная матрица, сформированная *Пакетом анализа* и наблюдаемые и критические значения T – статистики

Так как $T_{\text{набл}} > t_{\text{kp}}$ – отвергаем нулевую гипотезу о равенстве нулю генерального коэффициента корреляции. Значит, X и Y линейно коррелированы.

Замечание 6.1. В пакете *Excel* для вычисления выборочных коэффициента корреляции и корреляционного момента можно использовать стандартные функции (см. рис. 6.7), которые находятся на вкладке *Формулы – Другие функции – Статистические*:

	A	B	C	D	E	F	G	H	I
1	**X**	0	1	2	4	6	8	9	10
2	**Y**	6	7,2	9,4	11	15,2	16,6	19,4	21,2
3									
4	**Коэффициент корреляции КОРРЕЛ(массив1; массив 2):**								**0,9939**
5									
6	**Корреляционный момент КОВАРИАЦИЯ.В(массив 1; массив 2):**								**21,514**
7									

Рис. 6.7 – Расчет коэффициента корреляции и корреляционного момента с помощью стандартных функций пакета *Excel*.

Контрольный пример 6.2. Два преподавателя оценили знания 12 учащихся по стобалльной системе и выставили им следующие оценки (в первой строке указано количество баллов, выставленных первым преподавателем, а во второй – вторым):

Студент	1	2	3	4	5	6	7	8	9	10	11	12
Преподаватель 1 (*X*)	88	94	98	80	76	63	56	60	58	66	51	61
Преподаватель 2 (*Y*)	93	91	99	74	78	64	48	52	53	68	62	66

Используя коэффициенты ранговой корреляции Спирмена и Кендалла, проверить на уровне значимости $\alpha = 0{,}05$ гипотезу о полной несогласованности (независимости) оценок преподавателей против альтернативной – оценки экспертов находятся в согласии (зависимы).

Задание выполнить в пакетах Statistica и Excel.

Решение.

1) Для нахождения коэффициента ранговой корреляции Спирмена в пакете *Statistica* создаем новый документ, в графе *Number of cases* (количество значений) пишем 10. После чего вводим значения переменных (рис.6.8).

	1 X	2 Y
1	88	93
2	94	91
3	98	99
4	80	74
5	76	78
6	63	64
7	56	48
8	60	52
9	58	53
10	66	68
11	51	62
12	61	66

Рис.6.8 – Ввод значений переменных

После этого, открываем меню *Statistics* и выбираем пункт *Nonparametrics* (непараметрические статистики).

В появившемся меню открываем пункт *Correlation*. В результате открывается окно для настроек расчета коэффициента корреляции Спирмена.

Нажимаем кнопку *Two lists*, выбираем переменные, для которых нужно установить корреляционную зависимость, после чего возвращаемся в главное меню настроек (рис 6.9).

В пункте *Compute* выбираем один из способов вывода результата: *Detailed report* (подробный отчет) или *Matrix or two lists* (вывод лишь подсчитанного коэффициента ранговой корреляции).

В пункте *p-level for highlighting* выбираем требующееся значение уровня значимости.

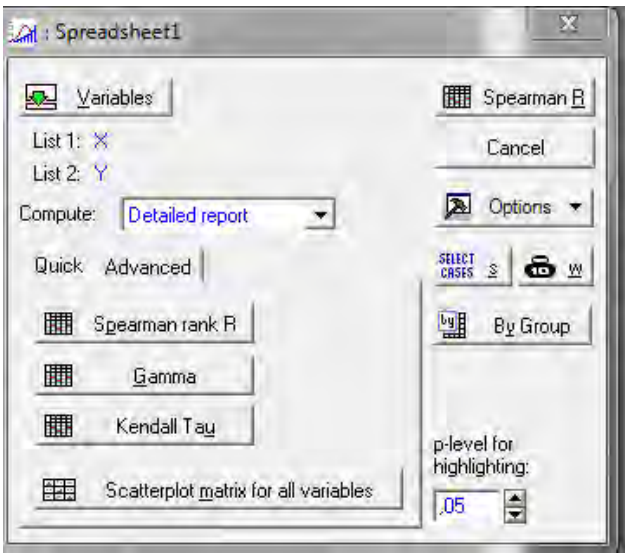


Рис.6.9 – Меню настроек для вычисления коэффициента Спирмена

Вывод результата можно осуществить двумя способами: нажав в правом верхнем углу окна кнопку *Spearman R (Kendall tau)*, либо, открыв одну из вкладок *Advanced* или *Quick*, нажать одноименную кнопку. Вне зависимости от выбора способа, на экране возникнет вычисленный результат (рис.6.10).

Spearman Rank Order Correlations (Spreadsheet1)
MD pairwise deleted
Marked correlations are significant at p <,05000

Pair of Variables	Valid N	Spearman R	t(N-2)	p-level
X & Y	12	0,930070	8,005659	0,000012

Рис.6.10 – Коэффициент ранговой корреляции Спирмена

Коэффициент корреляции Спирмена равен 0,93.

Гипотеза о полной несогласованности оценок преподавателей противоречит данным наблюдения, поэтому её нужно отклонить на фактическом уровне значимости $p = 0{,}000012$, который меньше номинального уровня значимости $\alpha = 0{,}05$. Этот вывод подтверждается и с помощью коэффициента ранговой корреляции Кендалла (рис. 6.11).

Kendall Tau Correlations (Spreadsheet1)
MD pairwise deleted
Marked correlations are significant at p <,05000

Pair of Variables	Valid N	Kendall Tau	Z	p-level	p-exact 1-tailed
X & Y	12	0,787879	3,565773	0,000363	----

Рис.6.11 – Коэффициент ранговой корреляции Кендалла

Здесь $p = 0{,}00063 < 0{,}05$.

Рассмотренные примеры отличаются малым числом наблюдений. Для надёжного результата общее число наблюдений не должно быть меньше 50.

Несоблюдение этого требования не гарантирует достаточно точных выводов, которые делают на основании выборочных показателей.

2) Решим эту задачу, используя инструменты *MS Excel*. Для этого:

Сформируем таблицу исходных данных (рис. 6.12).

	A	B	C	D	E
1	**Студент**	**Преподаватель 1**	**Ранги ri**	**Преподаватель 2**	**Ранги si**
2	1	88		93	
3	2	94		91	
4	3	98		99	
5	4	80		74	
6	5	76		78	
7	6	63		64	
8	7	56		48	
9	8	60		52	
10	9	58		53	
11	10	66		68	
12	11	51		62	
13	12	61		66	

Рис. 6.12 – Исходные данные

Выберем ячейку C2 и перейдем на вкладку *Формулы – Другие функции – Статистические* и в раскрывающемся списке выберем функцию **РАНГ.СР**. В появившемся окне введем данные: **Число** – B2, **Ссылка** – $B$2:$B$13 (диапазон ранжируемых значений), **Порядок** – 1 (рис. 6.13).

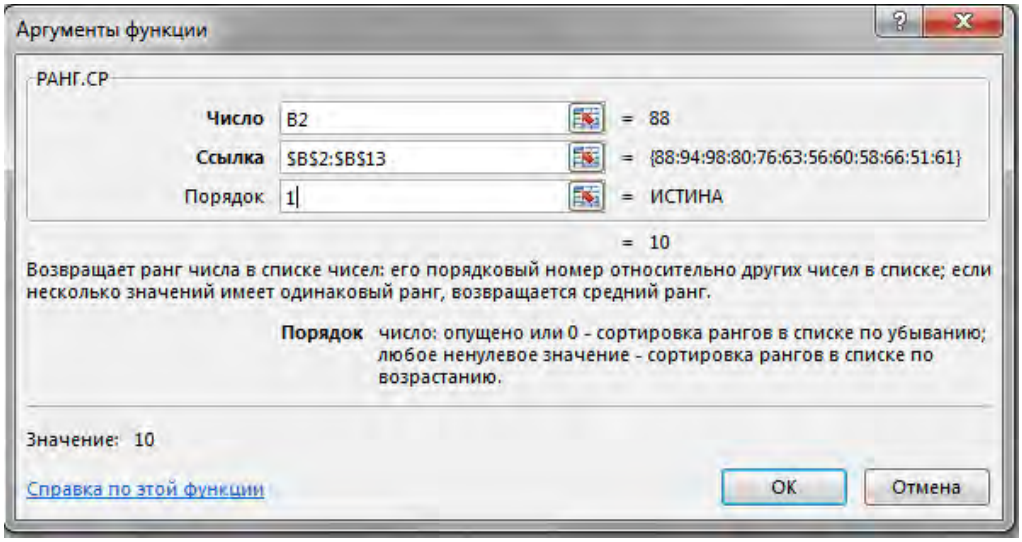


Рис. 6.13 – Диалоговое окно функции РАНГ.СР

Снова выберем ячейку C2 и растянем формулы по столбцу. Все оценки первого эксперта будут проранжированы. То же самое проделаем со столбцом **Ранг si**.

Теперь найдем разности рангов. В ячейку F2 введем функцию =C2–E2 и перенесем формулы по столбцу до последнего значения.

Посчитаем квадраты разности рангов. Для этого выберем ячейку G2 и перейдем на вкладку *Формулы – Математические* и в раскрывающемся списке выберем функцию СТЕПЕНЬ. В появившемся окне в поле **Число** вводим F2, в поле **Степень** – 2 и переносим формулы по столбцу до последнего значения (рис. 6.14).

Выделим диапазон G2:G13 и, перейдя на вкладку *Формулы*, выберем инструмент ***Автосумма***.

Замечание 6.2. В ячейках C14 и E14 записаны суммы рангов оценок первого и второго экспертов, при отсутствии совпадений рангов эти суммы должны равняться числу $\frac{n(n+1)}{2}$; в данном примере $\frac{12\cdot(12+1)}{2}=78$.

Замечание 6.3. В ячейке G15 записана выборочная оценка коэффициента ранговой корреляции Спирмена, вычисленная по формуле:

$$=1 – 6*G14/(12^3 – 12).$$

	A	B	C	D	E	F	G	H	I	J
1	Студент	Преподаватель 1	Ранги ri	Преподаватель 2	Ранги si	Разности рангов	(ri-si)^2			
2	1	88	10	93	11	-1	1		t =	8,00565923
3	2	94	11	91	10	1	1		t(0,025;8)=	2,22813885
4	3	98	12	99	12	0	0		p =	1,1702E-05
5	4	80	9	74	8	1	1			
6	5	76	8	78	9	-1	1			
7	6	63	6	64	5	1	1			
8	7	56	2	48	1	1	1			
9	8	60	4	52	2	2	4			
10	9	58	3	53	3	0	0			
11	10	66	7	68	7	0	0			
12	11	51	1	62	4	-3	9			
13	12	61	5	66	6	-1	1			
14			78		78	S =	20			
15						ρB =	0,93007			

Рис. 6.14 – Расчёт коэффициента ранговой корреляции Спирмена

Замечание 6.4. В ячейках J2:J4 находятся наблюдаемое значение *t* статистики *T*, её критическое значение (критическое значение распределения Стьюдента) и значимость *p* (см. рис. 6.15):

I	J
t =	=G15*КОРЕНЬ(10/(1-G15^2))
t(0,025;8)=	=СТЬЮДЕНТ.ОБР.2Х(0,05;10)
p =	=СТЬЮДЕНТ.РАСП.2Х(J2;10)

Рис. 6.15 – Формулы для расчёта наблюдаемого значения статистики *T*, её критического значения и значимости *p*

Так как $|t| > t(0{,}025;10)$ и $p < \alpha$ – гипотеза о полной несогласованности оценок преподавателей противоречит данным наблюдения и её надо отклонить.

Теперь проверим с помощью рангового критерия независимости Кендалла гипотезу о несогласованности (независимости) оценок преподавателей.

Для этого:

1. Откроем новый рабочий лист и скопируем на него оценки преподавателей и ранги этих оценок (рис. 6.16, диапазон A1:E13).

2. Выделим диапазон, в котором находятся ранги оценок и щелкнем на кнопке «*Копировать*».

3. Выделим ячейку F1, выберем в меню *Главная* – *Вставить* – *Специальная вставка* – *Значения*. В диапазоне F2:G13 появятся «копии» рангов экспертных оценок.

	A	B	C	D	E	F	G	H	I
1	**Студент**	**Преподаватель 1**	**Преподаватель 2**	**Ранги ri**	**Ранги si**	**ri**	**si**	**Ri**	
2	1	88	93	10	11	1	4	8	
3	2	94	91	11	10	2	1	10	
4	3	98	99	12	12	3	3	8	
5	4	80	74	9	8	4	2	8	
6	5	76	78	8	9	5	6	6	
7	6	63	64	6	5	6	5	6	
8	7	56	48	2	1	7	7	5	
9	8	60	52	4	2	8	9	3	
10	9	58	53	3	3	9	8	3	
11	10	66	68	7	7	10	11	1	
12	11	51	62	1	4	11	10	1	
13	12	61	66	5	6	12	12		
14							**R =**	**59**	
15							**τB =**	**0,788**	

Рис. 6.16 – Исходные данные и решение контрольного примера 6.2 (коэффициент Кендалла)

4. Выделим диапазон F1:G13. Выбираем *Сортировка и фильтр* – *настраиваемая сортировка*. В открывшемся окне в столбце *Сортировать по* выберем поле *ri*, по которому надо выполнить сортировку, установим порядок – *по возрастанию* и щёлкнем на кнопке ОК.

В диапазоне F2:G13 появятся ранги оценок преподавателей, отсортированные в порядке возрастания рангов оценок первого эксперта.

5. В ячейку H2 введём формулу массива =СУММ(ЕСЛИ($G3:$G$13>G2; 1; 0)), нажмём клавиши CTRL – SHIFT – ENTER (одновременно) и затем скопируем эту формулу в ячейки H3:H12. В диапазоне H2:H12 появятся числа $R_1 = 8, R_2 = 10, \ldots, R_{11} = 1$.

6. Суммируя эти числа, находим $R = 59$ (ячейка H14).

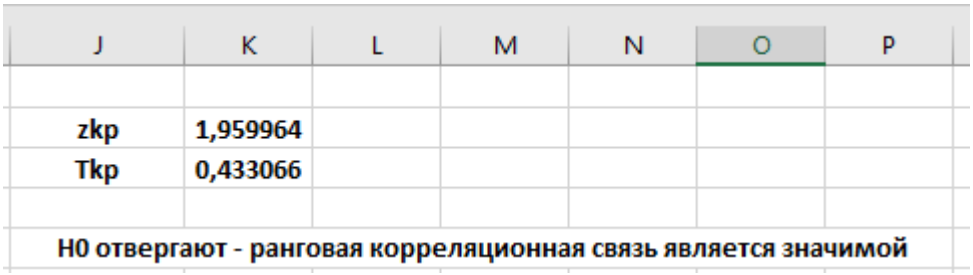
7. Используя формулу =4*H14/12/11 – 1, находим выборочный коэффициент ранговой корреляции Кендалла $\tau_B \approx 0{,}788$.

Проверим коэффициент ранговой корреляции Кендалла на значимость.

1) Найдем критическую точку z_{kp} с помощью стандартной функции пакета Excel =НОРМ.ОБР(1 – 0,05/2; 0; 1) (рис. 6.17, ячейка К2).

2) Найдем критическую точку (рис. 6.17, ячейка К3) по формуле

$$T_{\text{kp}} = z_{\text{kp}}\sqrt{\frac{2(2n+5)}{9n(n-1)}} = 1{,}96\cdot\sqrt{\frac{2\cdot(2\cdot 12+5)}{9\cdot 2\cdot 11}} = 0{,}433066.$$



J	K	L	M	N	O	P
zkp	1,959964					
Tkp	0,433066					
H0 отвергают - ранговая корреляционная связь является значимой						

Рис. 6.17 – Проверка коэффициента ранговой корреляции Кендалла на значимость

Так как $|\tau_B| > T_{\text{kp}}$ – ранговая корреляционная связь является значимой.

Таким образом, оба ранговых критерия (и Спирмена, и Кендалла) свидетельствуют о том, что гипотеза о полной несогласованности (независимости) мнений преподавателей противоречит данным наблюдения.

Контрольный пример 6.3**.** В табл. 6.2 приведены данные, полученные в результате эксперимента, целью которого являлось определение тесноты связи между объемом выпуска продукции и температурой определенного технологического процесса.

Требуется:

- Построить диаграмму рассеяния (корреляционное поле) для этой совокупности данных (в пакете *Statistica*).
- Оценить тесноту связи между объемом выпуска продукции и температурой (в пакете *Excel*).

Таблица 6.2

Температура (x)	600	625	650	675	700	725	750	775	800	825	850
Объем выпуска продукции (Y)	127	139	147	147	155	154	153	148	146	136	129

*Решение***.** Введём исходные данные в пакете *Statistica* (рис. 6.18):

Data: Spreadsheet...

	1 x	2 Y
1	600	127
2	625	139
3	650	147
4	675	147
5	700	155
6	725	154
7	750	153
8	775	148
9	800	146
10	825	136
11	850	129

Рис. 6.18 – Исходные данные

Для построения поля корреляции открываем меню *Graphs* и выбираем вкладку *Scatterplot – Advanced – Graph type/Regular – Fit/Off* ; нажмём кнопку OK, после чего на экране появляется график корреляционного поля (рис. 6.19):

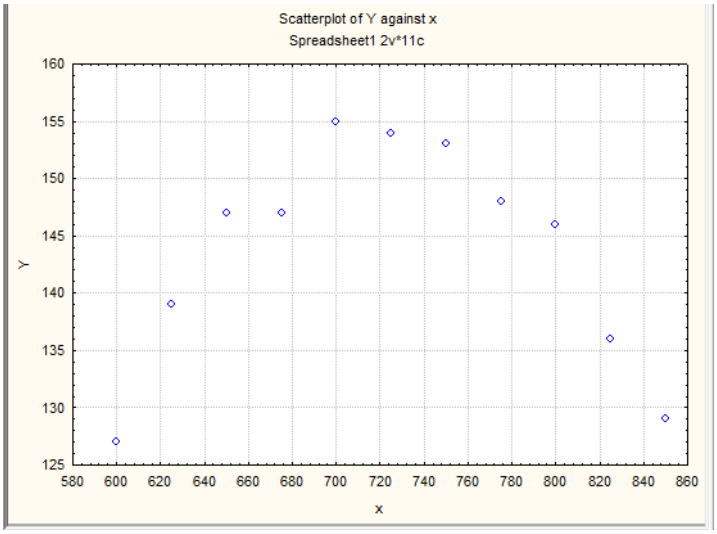


Рис. 6.19 – Диаграмма рассеивания

Корреляционное поле, показанное на рис. 6.19, иллюстрирует сильную нелинейную взаимосвязь, характеризующуюся незначительным случайным разбросом.

Также корреляционное поле демонстрирует, что для максимального увеличения объема выпускаемой продукции температуру производственного процесса следует установить равной примерно 700 градусам. Объем продукции резко падает как при слишком высокой, так и при слишком низкой температуре. Этот важный вывод можно сделать, наблюдая на диаграмме сильную взаимосвязь между объемом продукции и температурой.

Найдем значение коэффициента корреляции (рис. 6.20), используя пункт *Correlation matrices* меню *Statistics*:

Var. X & Var. Y	Correlations (Spreadsheet1) Marked correlations are significant at p < ,05000 (Casewise deletion of missing data) **Mean**	Std.Dv.	r(X,Y)	r?	t	p	N
x	725,0000	82,91562					
Y	143,7273	9,70661	-0,015531	0,000241	-0,046599	0,963850	11

Рис. 6.20 – Вычисление коэффициента корреляции и основных выборочных характеристик величин x и Y

Выборочный коэффициент парной корреляции $r_B = -0{,}0155$ бесполезен в случае такой нелинейной связи: он не может решить, является связь увеличивающей или уменьшающей, поскольку в действительности есть и то и другое.

Замечание 6.5*:* Близкое к нулю значение коэффициента корреляции может означать как отсутствие взаимосвязи в данных, так и наличие нелинейной взаимосвязи без преобладания направленности вниз или вверх. Сильная нелинейная взаимосвязь может быть даже тогда, когда корреляция близка к нулю.

Работаем в пакете *Excel.*

Оценим тесноту связи между объемом выпуска продукции и температурой с помощью корреляционного отношения $\eta_{Y/x}$.

Введем исходные данные на лист *MS Excel*, как показано на рис. 6.21. Значения результирующего признака Y разобьем на 5 групп $(l = 5)$. В основу группировки кладётся исследуемый фактор x:

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Температура (x)**	600	625	650	675	700	725	750	775	800	825	850
2	**Объем выпуска продукции (Y)**	127	139	147	147	155	154	153	148	146	136	129
3												
4	**Коэффициент корреляции rв**	**-0,01553**		$\bar{y}$	**143,727**							
5												
6	**Номер группы**	**1**	**2**	**3**	**4**	**5**		**Общая дисперсия**			**85,653**	
7	**Значения y, попавшие в i-ю группу**	127	147	155	148	136		**Межгрупповая дисперсия**			**76,517**	
8		139	147	154	146	129		**Корреляционное отношение** $\eta_{Y/x}$			**0,945**	
9				153								
10	**Количество элементов выборки в i-ой группе**	2	2	3	2	2						
11	**Среднее значение y в i-й группе** $(\bar{y}_{xi})$	133	147	154	147	132,5						
12	$(\bar{y}_{xi} - \bar{y})^2$	115,07	10,71	105,53	10,71	126,05						
13												

Рис. 6.22 – Вид листа *MS Excel* с исходными данными и расчетами для вычисления корреляционного отношения

Для каждой группы рассчитаем *групповую среднюю*.

В ячейку B11 введем формулу =СРЗНАЧ(B7:B9), которую затем скопируем методом автозаполнения вправо по строке. При этом *Excel* игнорирует пустые ячейки при расчете среднего значения (как и при использовании других статистических функций). Общую среднюю рассчитаем с помощью функции =СРЗНАЧ(B2:L2) (ячейка E4)

В ячейке B4 вычислим выборочный коэффициент корреляции по формуле =КОРРЕЛ(B1:L1; B2:L2).

Рассчитаем квадраты отклонений групповых средних от общей средней (диапазон B12:E12). Введем в ячейку B12 формулу =(B11 – $E$4)^2, которую затем скопируем вправо по строке.

Межгрупповую дисперсию рассчитаем в ячейке K7 по формуле = СУММПРОИЗВ(B12:F12; B10:F10)/11, а в ячейке K6 находится значение общей дисперсии, вычисленной с помощью стандартной функции пакета ДИСП.Г(B2:L2).

Корреляционное отношение $\eta_{Y/x} = \sqrt{\frac{76{,}517}{85{,}653}} \approx 0{,}95$ рассчитано в ячейке K7 по формуле =КОРЕНЬ(K7/K6). Полученное значение свидетельствует о наличии сильного нелинейного влияния температуры на объем выпуска продукции.

Контрольный пример 6.4. Экспертная комиссия из 5 человек проранжировала 7 сочинений школьников – участников олимпиады по математике (ранг 1 присваивался самой лучшей работе). Ранговые последовательности приведены в таблице 6.3.

Таблица 6.3

Школьники (n)	Эксперты (m)				
	1	**2**	**3**	**4**	**5**
1	1	1	2	1	3
2	3	2	1	2	1
3	4	5	7	4	5
4	2	3	5	6	4
5	6	6	6	3	2
6	7	4	4	5	6
7	5	7	3	7	7

Требуется вычислить коэффициент конкордации. Проверить гипотезу $H_0 : W = 0$ о несогласованности (независимости) экспертных оценок. Принять $\alpha = 0{,}05$.

Решение. В расчетную таблицу 6.4 заносим экспертные оценки, ранговые суммы $d_i = \sum_{j=1}^{m} R_{ij}$, отклонения $D_i = d_i - \bar{d}$ суммы рангов от средней $\bar{d} = \frac{\sum d_i}{7}$ и D_i^2.

Средняя сумма рангов всех объектов равна: $\bar{d} = \frac{140}{7} = 20$.

В качестве контроля используем выражение: $\bar{d} = \frac{1}{2} m(n+1) = \frac{5 \cdot 8}{2} = 20$.

Таблица 6.4

Школьники (n)	**Эксперты** (m)					d_i	D_i	D_i^2
	1	**2**	**3**	**4**	**5**			
1	1	1	2	1	3	8	–12	144
2	3	2	1	2	1	9	–11	121
3	4	5	7	4	5	25	5	25
4	2	3	5	6	4	20	0	0
5	6	6	6	3	2	23	3	9
6	7	4	4	5	6	26	6	36
7	5	7	3	7	7	29	9	81
Сумма						**140**		**416**

Коэффициент конкордации Кендалла определяется по формуле:

$$W = \frac{12 \cdot 416}{5^2 \left(7^3 - 7\right)} = 0{,}594.$$

Проверку нулевой гипотезы о том, что мнения экспертов не согласуются друг с другом ($H_0 : W = 0$) проводим с помощью критерия Пирсона «хи–квадрат». Для этого вычисляем эмпирическое значение

$$\chi^2 = 5 \cdot 6 \cdot 0{,}594 = 17{,}82,$$

которое сравниваем с критическими значениями χ^2 для числа степеней свободы $k = n - 1 = 6$: $\chi^2_{\text{кр}}(0{,}05; 6) = 12{,}592$.

Эмпирическое значение $\chi^2 = 17{,}82$ попадает в критическую область $\left(\chi^2 > \chi^2_{\text{кр}}\right)$, что позволяет отвергнуть нулевую гипотезу. Коэффициент конкордации значимо отличается от нуля, следовательно имеется достаточно тесная согласованность мнений экспертов относительно сочинений.

Задания для самостоятельной работы

Задание 1. Для исследования зависимости случайных величин X и Y получены статистические данные.

Требуется:

1) Построить корреляционное поле, сделать вывод.
2) Найти корреляционный момент и выборочный коэффициент корреляции.
3) При уровне значимости $\alpha = 0{,}05$ проверить нулевую гипотезу о равенстве генерального коэффициента корреляции нулю при конкурирующей гипотезе $H_1 : r_{XY} \neq 0$.

Задание выполнить в пакетах *Statistica*, *Excel***.**

1	X	0	1	2	3	4	5	6	7	8	9
	Y	1	1,4	2,5	2,5	3	3,5	4,1	4,5	5	5,5
2	X	–1	–0,7	–0,4	–0,2	0,1	0,4	0,7	1	1,2	1,5
	Y	–5,4	–5	–4,5	–4	–3	–1,9	–1	0,1	0,3	0,6
3	X	1	2	3	4	5	6	7	8	9	10
	Y	0,3	1	1,1	1,8	2,1	3	3,2	3,7	4,2	4,5
4	X	0,4	0,8	1,3	1,8	2,2	2,7	3,1	3,6	4,1	4,5
	Y	5,7	5,2	5	4,6	4,5	4,4	4,35	4,2	4	3,5
5	X	0,3	0,91	1,5	2	2,2	2,62	3	3,3	3,5	
	Y	0,2	0,42	0,48	0,6	0,7	0,81	1	0,9	1,04	
6	X	0	1	2	3	4	5	6	7	8	9
	Y	–4	–3,2	–2	–1	–0,7	0	0,5	1	1,6	2,1
7	X	0	0,4	0,8	1,2	1,6	2	2,4	2,8	3,2	3,6
	Y	5	7,5	11,4	14,5	17,2	19,9	22	24,1	26,1	28
8	X	-3,5	-2,7	-1,8	-1	-0,1	0,8	1,6	2,5	3,3	4,2
	Y	0,01	0,03	0,07	0,12	0,19	0,21	0,24	0,28	0,33	0,35
9	X	0,1	0,4	0,7	1	1,3	1,6	1,9	2,2	2,5	2,8
	Y	1,5	1,3	1,2	1,1	1,03	0,9	0,7	0,6	0,4	0,25
10	X	0	1	3	6	8	10	11	13	15	19
	Y	3,2	4,3	5,4	8,3	9	11,4	11,7	13,8	15,1	18,2
11	X	-2	-1	0	1	2	3	4	5	6	7
	Y	-0,4	0,2	0,7	1,6	2	3,5	4	4,2	4,8	5,2
12	X	0,4	0,8	1,3	1,8	2,2	2,7	3,1	3,6	4,1	4,5
	Y	5,7	5,2	5	4,6	4,5	4,4	4,35	4,2	4	3,5

Задание 2. Два преподавателя (А и В) оценили знания нескольких учащихся по стобальной системе и выставили им следующие оценки (в первой строке указано количество баллов, выставленных первым преподавателем, а во второй – вторым).

Найти значение ранговых коэффициентов корреляции Спирмена и Кендалла, провести анализ результатов; проверить их на значимость, приняв $\alpha = 0{,}05$.

Задание выполнить в пакетах *Statistica*, *Excel***.**

1	A	95	91	90	88	85	86	71	70	68	65
	B	96	89	91	86	84	71	72	69	56	70
2	A	94	92	91	81	80	74	73	72	62	60
	B	84	80	88	91	71	79	77	83	63	66
3	A	95	91	90	83	76	75	71	70	65	61
	B	96	90	87	84	66	67	74	68	79	64
4	A	99	98	97	96	88	82	81	77	74	73
	B	83	88	90	89	81	85	79	72	82	75
5	A	89	85	83	81	80	76	74	71	68	59
	B	78	79	81	86	77	73	70	65	90	61
6	A	98	91	90	87	86	80	75	72	69	61
	B	95	78	91	70	85	81	88	69	59	68
7	A	91	90	85	83	81	79	74	71	69	65
	B	87	93	81	92	89	80	73	69	61	83
8	A	87	85	82	81	76	74	65	61	58	54
	B	86	78	81	84	80	77	76	59	61	56
9	A	99	96	94	88	87	84	83	81	79	77
	B	93	82	87	95	88	91	89	90	81	80
10	A	98	94	91	87	85	84	83	81	80	76
	B	89	90	93	86	92	77	79	82	84	74
11	A	97	95	94	90	84	82	80	76	72	71
	B	89	85	86	96	87	95	90	79	78	88
12	A	99	91	86	84	83	76	75	74	72	70
	B	88	93	87	91	86	90	74	77	80	82

Задание 3. Ниже приведены данные, полученные в результате эксперимента, целью которого являлось определение тесноты связи между признаками X и Y.

Требуется:

- Построить диаграмму рассеяния (корреляционное поле) для этой совокупности данных (в пакете *Statistica*).
- Оценить тесноту связи между данными признаками (в пакете *Excel*).

Вариант 1

X	–0,6	–0,3	0	0,3	0,6	0,9	1,2	1,5
Y	11,3	9,4	8,1	6,6	5,6	4,4	3,3	2,8
X	1,8	2,1	2,4	2,7	3	3,3	3,6	3,9
Y	2,3	1,9	1,3	1,8	2,2	2,5	3	4,7

Вариант 2

X	–0,4	–0,2	0	0,2	0,4	0,6	0,8	1
Y	5,2	4,6	3,8	3,6	3,4	3,5	3,8	4
X	1,2	1,4	1,6	1,8	2	2,2	2,4	2,6
Y	4,5	5,2	5,8	7,1	8,2	9,3	10,8	12,3

Вариант 3

X	–0,5	–0,3	–0,1	0,1	0,3	0,5	0,7	0,9
Y	–0,1	–0,2	–0,8	–1,2	–1,4	–1,5	–1,3	–1,1
X	1,1	1,3	1,5	1,7	1,9	2,1	2,3	2,5
Y	–0,7	–0,2	0,4	1,4	2,4	3,6	5,1	6,6

Вариант 4

X	0,1	0,5	0,9	1,3	1,7	2,1	2,4	2,8
Y	1,3	0,6	–0,1	–0,37	–0,52	–0,41	0	0,52
X	3,2	3,6	4	4,4	4,8	5,2	5,6	5,8
Y	1,55	2,2	2,8	4,2	5,8	7,5	9,6	11,7

Вариант 5

X	0	0,3	0,6	0,9	1,2	1,5	1,8	2,1
Y	9,3	7,4	6,4	4,8	3,6	2,7	1,8	1,2
X	2,4	2,7	3	3,3	3,6	3,9	4,2	4,5
Y	0,8	0,6	0,5	0,3	0,7	1,3	2,1	3

Вариант 6

X	–1	–0,8	–0,6	–0,4	–0,2	0	0,2	0,4
Y	–0,1	–0,5	–1,1	–1,6	–1,9	–1,3	–1,2	–0,8
X	0,6	0,8	1	1,2	1,4	1,6	1,8	2
Y	1,3	2,6	4,2	5,7	7,7	8,4	9,9	10,3

Вариант 7

X	0,7	0,8	0,9	1	1,1	1,2	1,3	1,4
Y	3,1	4,9	5,3	5,8	6,1	6,1	5,9	5,5
X	1,5	1,6	1,7	1,8	1,9	2	2,1	2,2
Y	4,2	2,2	0,8	–1,3	–3,7	–6,5	–9,6	–10,2

Вариант 8

X	–0,6	–0,3	0	0,3	0,6	0,9	1,2	1,5
Y	9,4	7,6	6	4,6	3,4	2,3	1,4	0,8
X	1,8	2,1	2,4	2,7	3	3,3	3,6	4
Y	0,3	–0,1	–0,3	–0,2	0,1	0,4	1	1,5

Вариант 9

X	-0,6	-0,3	0	0,3	0,6	0,9	1,2	1,5
Y	9,8	7,9	6,4	5,1	3,9	2,7	1,8	1,2
X	1,8	2,1	2,4	2,7	3	3,3	3,6	4
Y	0,7	0,3	0,1	0,2	0,5	0,8	1,4	2

Вариант 10

X	–0,5	-0,3	-0,1	0,1	0,3	0,5	0,7	0,9
Y	0,9	0,7	0,2	-0,2	-0,5	-0,6	-0,4	-0,2
X	1,1	1,3	1,5	1,7	1,9	2,1	2,3	2,5
Y	0,3	0,7	1,5	2,3	3,5	4,7	6,2	7,7

Вариант 11

X	0	0,3	0,6	0,9	1,2	1,5	1,8	2,1
Y	10,2	8,3	7,4	4,7	4,5	3,6	2,9	2,4
X	2,4	2,7	3	3,3	3,6	3,9	4,2	4,5
Y	1,8	1,5	1,4	1,2	1,8	2,4	3,3	3,8

Вариант 12

X	–2	–1,2	–0,4	0,4	1,2	2	2,8	3,6
Y	2	1,5	1,2	0,9	0,7	0,8	1	1,5
X	4,4	5,2	6	6,8	7,6	8,4	9,2	10
Y	2	2,5	2,8	3,6	4,5	5,5	6,6	7,7

Задание 4. Необходимо, используя коэффициент конкордации**,** определить степень согласованности мнений экспертов. Принять $\alpha = 0{,}05$.

Вариант 1. Тремя экспертами было оценено предприятие по 10 показателям. В конечном итоге получены три последовательности рангов:

Первый эксперт	1	2	3	6	5	4	7	8	9	10
Второй эксперт	3	10	7	2	5	8	6	9	1	4
Третий эксперт	6	1	2	3	9	4	7	5	10	8

Вариант 2. Три опытных специалиста оценили предприятие по 10 показателям и получили три последовательности рангов.

Первый эксперт	2	1	3	4	5	6	7	8	10	9
Второй эксперт	3	7	10	2	8	5	6	9	1	4
Третий эксперт	6	2	1	3	9	4	5	7	8	10

Вариант 3. Экспертами Ивановым С.И., Кудряшовым С.С. и Лемеховым Д.А. было оценено предприятие по 10 показателям. Ими были получены три последовательности рангов:

Иванов С.И.	1	2	3	4	5	6	7	8	9	10
Кудряшов С.С.	3	9	7	2	8	5	10	6	1	4
Лемехов Д.А.	6	10	1	3	9	4	5	7	2	8

Вариант 4. Три эксперта оценили предприятие по 10 показателям. Были получены три последовательности рангов:

Первый эксперт	1	2	3	6	5	4	7	8	9	10
Второй эксперт	3	10	7	2	8	5	6	9	1	4
Третий эксперт	6	2	1	3	9	4	5	7	10	8

Вариант 5. Специалисты трех заводов Столяров А.П., Иванов С.И и Хлеборезов С.В. проранжировали 11 факторов, влияющих на ход технологического процесса очистки сточных вод. В результате были получены три последовательности рангов.

Столяров А.П.	1	2	3	4	5	6	7	8	9	10	11
Иванов С.И.	3	1	2	5	4	11	8	9	6	7	10
Хлеборезов С.В.	2	3	1	4	5	6	9	11	7	10	8

Вариант 6.

Три эксперта оценили предприятие по 10 показателям. Были получены три последовательности рангов:

Первый эксперт	3	2	1	6	5	4	9	10	8	7
Второй эксперт	1	2	4	5	6	3	8	9	7	10
Третий эксперт	2	1	3	4	7	5	6	10	8	9

Вариант 7. Три эксперта оценили предприятие по 10 показателям. Были получены три последовательности рангов:

Первый эксперт	2	3	6	1	9	7	4	10	8	5
Второй эксперт	1	3	7	2	10	6	5	8	9	4
Третий эксперт	4	2	5	1	8	9	6	10	7	3

Вариант 8. Эксперты Соколов А.В., Гипнов С.З., Пластинин А.С. оценили предприятие по 10 показателям и получили три последовательности рангов.

Соколов А.В	1	2	3	6	5	4	7	8	9	10
Гипнов С.З	3	10	7	2	5	8	6	9	1	4
Пластинин А.С	6	1	2	3	9	4	7	5	10	8

Вариант 9.Три эксперта оценили предприятие по 10 показателям. Были получены три последовательности рангов:

Первый эксперт	1	5	4	3	2	8	9	7	10	6
Второй эксперт	1	5	2	6	3	4	7	10	9	8
Третий эксперт	2	3	4	5	1	6	8	9	10	7

Вариант 10. Специалистами трех заводов было проранжировано 11 факторов, влияющих на ход технологического процесса очистки сточных вод. В результате были получены три последовательности рангов.

1 специалист	1	2	3	4	5	6	7	8	9	10	11
2 специалист	1	2	3	5	4	7	8	11	6	9	10
3 специалист	2	1	3	4	5	6	7	8	10	11	9

Вариант 11. Три специалиста (В.И. Сухих, С.М. Артамонов и Б.Ф. Лапин) из разных лабораторий расположили пробы воды в порядке убывания содержания в них фенола. В итоге были получены три последовательности рангов:

Сухих В.И.	1	2	3	6	5	4	7	8	9
Лапин Б.Ф.	4	1	6	3	2	5	9	6	7
Артамонов С.М.	3	1	2	6	9	8	5	7	6

Вариант 12. Три специалиста (А.А. Листов, В.И. Сухих и Г.С. Туркин) из разных лабораторий расположили пробы воды в порядке убывания концентрации фенола. Получены три последовательности рангов:

Листов А.А.	1	2	3	4	5	6	7	8	9
Сухих В.И	4	1	5	3	2	6	9	8	7
Туркин Г.С.	2	1	3	5	4	8	9	7	6

Лабораторная работа № 7. Регрессионный анализ

Необходимые теоретические сведения для выполнения лабораторной работы находятся в теоретическом разделе – тема 4.2.

Контрольный пример 7.1. При исследовании зависимости между случайными величинами X и Y была получена следующая таблица измерений соответствующих значений этих величин (см. лабораторную работу № 6):

X	0	1	2	4	6	8	9	10
Y	6	7,2	9,4	11	15,2	16,6	19,4	21,2

Требуется (в пакетах *Statistica* и *Excel*):

1. Аппроксимировать статистическую зависимость между этими величинами линейной функцией $\bar{y}_x = a_0 + a_1 x$ проверить модель на значимость (адекватность).

2. Вычислить коэффициент детерминации, сделать вывод.
3. Построить корреляционное поле и линию регрессии на корреляционном поле. В пакете *Statistica* провести анализ остатков. Принять $\alpha = 0{,}05$.

Решение.

1. Выполнение в пакете *Statistica*.

Введем исходные данные (рис. 7.1).

	1 x	2 y
1	0	6
2	1	7,2
3	2	9,4
4	4	11
5	6	15,2
6	8	16,6
7	9	19,4
8	10	21,2

Рис. 7.1 – Исходные данные задачи

Будем работать в модуле *Multiple Regression* (множественная регрессия); меню *Statistics – Multiple Regression*. В качестве зависимой переменной выберем колонку *y,* в качестве независимой – колонку *x,* во вкладке *Advanced* установим опцию *Input file* (входной файл): *Raw Data* (необработанные данные). Нажав кнопку OK, получаем основные результаты анализа (рис. 7.2): имеем основные результаты: скорректированный коэффициент детерминации R^2 : 0,98578061; гипотеза о нулевом значении наклона отклоняется с высоким уровнем значимости $p = 0{,}000001$.

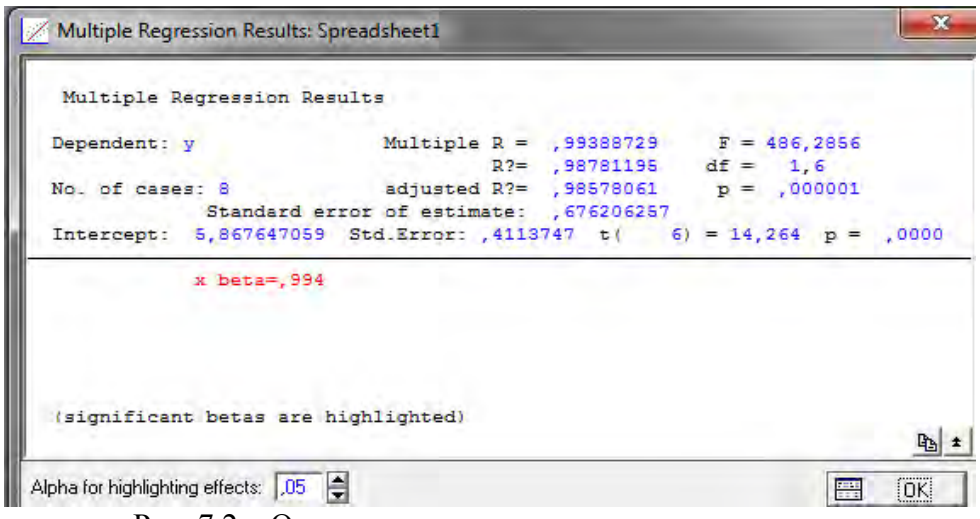


Рис. 7.2 – Окно результатов регрессионного анализа

Поясним значения характеристик:

- *Dependent* – имя зависимой переменной (в примере – *y*);
- *Multiple R* – множественный коэффициент корреляции (выборочный коэффициент корреляции);
- F – значение критерия Фишера, $F = 486{,}2856$;
- $R?(R^2)$ – множественный коэффициент детерминации;
- *df* – количество степеней свободы F-критерия;
- *No. of cases* – количество наблюдений;
- *adjusted* $R?(R^2)$ – скорректированный коэффициент детерминации;
- p – критический уровень значимости модели.
- *Standard error of estimate* – среднеквадратическая ошибка.
- *Intercept* – оценка свободного члена модели регрессии.
- *Std. Error* – стандартная ошибка оценки свободного члена модели регрессии.
- $t(6) = 14{,}264$ и $p = 0.0000$ – значения критерия и критического уровня значимости, используемые для проверки гипотезы о равенстве нулю свободного члена регрессии.

На вкладке *Quick* нажмем кнопку *Summary Regression Results* и получим таблицу результатов (см. рис. 7.3):

Regression Summary for Dependent Variable: y (Spreadsheet1)
R= ,99388729 R?= ,98781195 Adjusted R?= ,98578061
F(1,6)=486,29 p<,00000 Std.Error of estimate: ,67621

N=8	Beta	Std.Err. of Beta	B	Std.Err. of B	t(6)	p-level
Intercept			5,867647	0,411375	14,26351	0,000007
x	0,993887	0,045070	1,476471	0,066954	22,05188	0,000001

Рис. 7.3 – Таблица результатов регрессионного анализа

В заголовке полученной таблицы повторены результаты предыдущего окна; в столбцах приведены: *B* – значения оценок параметров модели регрессии $a_0 = 5{,}8676$ и $a_1 = 1{,}47647$; столбец *St. Err. of B* – параметры случайных ошибок параметров модели регрессии; столбец *t(6)* – значение статистики Стьюдента (*t*-критерия) для проверки гипотезы о нулевом значении коэффициента (т.е. $a_0 = 0$ и $a_1 = 0$); столбец *p-level* – минимальный уровень значимости отклонения этой гипотезы. В данном случае, поскольку значения *p-level* малы, гипотезы о нулевых значениях коэффициентов отклоняются с высокой значимостью.

Во второй вкладке *Summary Regression Results – Summary Statistics* – вычислены стандартная ошибка оценки по уравнению регрессии, коэффициент детерминации и наблюдаемое значение критерия Фишера (рис. 7.4):

Statistic	Summary Statistics; DV: y (Spreadsheet1) Value
Multiple R	0,9939
Multiple R?	0,9878
Adjusted R?	0,9858
F(1,6)	486,2856
p	0,0000
Std.Err. of Estimate	0,6762

Рис. 7.4 – Вкладка *Summary Statistics*

Так как стандартная ошибка оценки (*Std Error of estimate)* – небольшая, то наблюдаемые значения близки к предсказываемым. Значение коэффициента детерминации $R^{*2} = Adjusted\ R? = 0{,}9858$ достаточно велико.

Анализ остатков.

Для оценки адекватности модели необходимо исследовать остатки. *Остатки* – это разность между исходными (наблюдаемыми) значениями зависимой переменной и предсказанными (модельными, *Predicted values*) значениями. Остатки должны быть нормально распределены, иметь нулевое среднее значение и постоянную дисперсию, независимо от величин зависимых и независимых переменных. Модель должна быть адекватна на всех отрезках интервала изменения зависимой переменной. Вначале для оценки адекватности модели лучше всего использовать визуальные методы и затем, если потребуется, перейти к статистическим критериям.

В окне *Multiple Regression* выберем вкладку *Residuals/assumptions/prediction*, позволяющую оценить остатки и нажмем на кнопку *Perform Residual analysis*.

Далее кнопкой активизируем окно *Summary: Residual&Predicted* (рис. 7.5).

Первые четыре столбца этой таблицы определяют: номера наблюдений (названия областей), фактические (*Observed Value*) и расчетные значения (*Predicted Value*), отклонения фактических данных от расчетных (*Residual*).

Четыре последних строки содержат минимальное, максимальное, среднее и медианное значения показателей. Равенство нулю среднего значения остатков свидетельствует о корректности расчетов.

Выбросы – это остатки, которые значительно превосходят по абсолютной величине остальные. Выбросы дают данные, которые являются не типичными по отношению к остальным данным и требуют выяснения причин их возникновения.

Выбросы должны исключаться из обработки, если они вызваны ошибками измерения.

Case No.	Predicted & Residual Values (Spreadsheet1) Dependent variable: y								
	Observed Value	Predicted Value	Residual	Standard Pred. v.	Standard Residual	Std.Err. Pred.Val	Mahalanobis Distance	Deleted Residual	Cook's Distance
1	6,00000	5,86765	0,13235	-1,30984	0,19573	0,411375	1,715686	0,21012	0,017867
2	7,20000	7,34412	-0,14412	-1,04787	-0,21313	0,359003	1,098039	-0,20068	0,012413
3	9,40000	8,82059	0,57941	-0,78591	0,85686	0,312254	0,617647	0,73645	0,126461
4	11,00000	11,77353	-0,77353	-0,26197	-1,14393	0,248274	0,068627	-0,89405	0,117825
5	15,20000	14,72647	0,47353	0,26197	0,70027	0,248274	0,068627	0,54731	0,044155
6	16,60000	17,67941	-1,07941	0,78591	-1,59627	0,312254	0,617647	-1,37196	0,438889
7	19,40000	19,15588	0,24412	1,04787	0,36101	0,359003	1,098039	0,33993	0,035615
8	21,20000	20,63235	0,56765	1,30984	0,83946	0,411375	1,715686	0,90117	0,328655
Minimum	6,00000	5,86765	-1,07941	-1,30984	-1,59627	0,248274	0,068627	-1,37196	0,012413
Maximum	21,20000	20,63235	0,57941	1,30984	0,85686	0,411375	1,715686	0,90117	0,438889
Mean	13,25000	13,25000	0,00000	0,00000	0,00000	0,332726	0,875000	0,03353	0,140235
Median	13,10000	13,25000	0,18824	0,00000	0,27837	0,335629	0,857843	0,27502	0,080990

Рис. 7.5. Наблюдаемые и предсказанные значения остатков

Для выделения выбросов, имеющихся в регрессионных остатках, предложены следующие метрики:

1. *Расстояние Р. Д. Кука (Cook's Distance)* показывает расстояние между коэффициентами уравнения регрессии после исключения из обработки каждой точки данных. Большое значение показателя Кука указывает на сильно влияющее наблюдение. Так как в нашей таблице, приведённой на рис. 7.7 (последний столбец) больших значений нет – выбросы отсутствуют.

2. *Расстояние Махаланобиса (Mahalanobis Distance)* показывает, насколько каждое наблюдение отклоняется от центра статистической совокупности.

Построим диаграмму рассеяния и линию регрессии. Для этого в меню *Graphs* выберем команду *Scatterplots*. В полученном окне нажмем кнопку *Variables*, и установим зависимые данные – X, Y и опции графика – *Graphs Type: Regular, Fit (подбор)*: *Linear*. Наблюдаем диаграмму рассеяния с подобранной прямой регрессии, параметры которой отражены в ее заголовке (рис. 7.6).

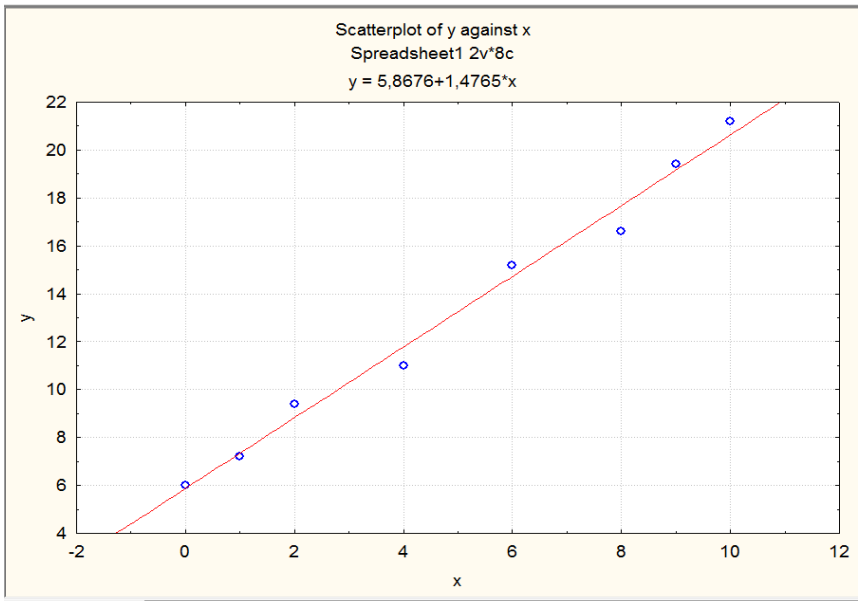
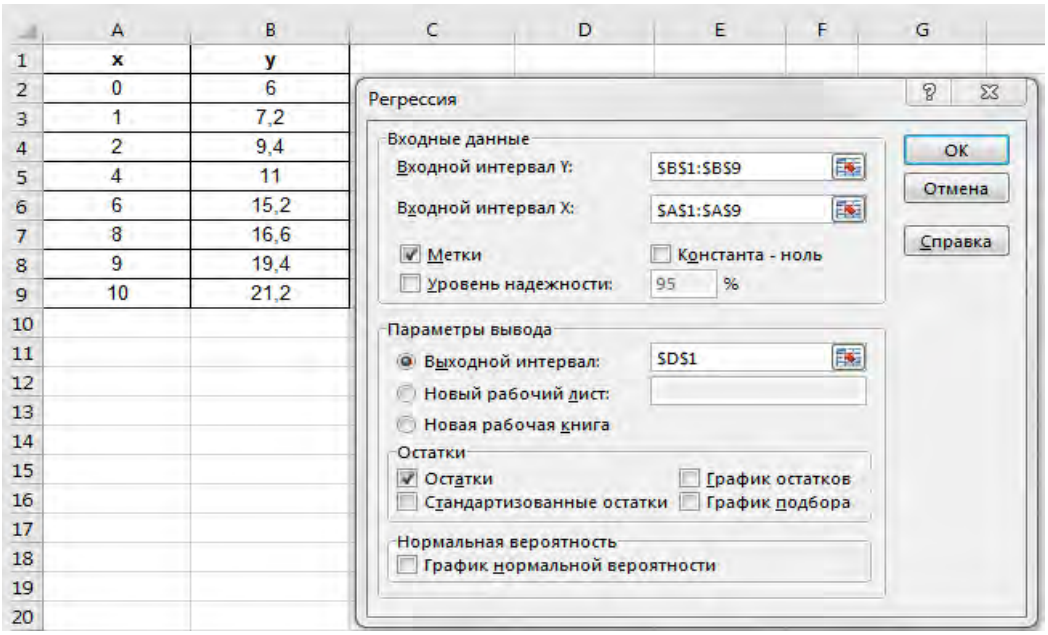


Рис. 7.6 – Диаграмма рассеяния с подобранной прямой линией регрессии

2) В пакете *Excel* построение линейной регрессии, оценивание ее параметров и их значимости выполним при помощи *надстройки* «*Пакет анализа*», которая находится на вкладке «*Данные*».

Введем исходные данные, расположив каждую случайную величину в отдельном столбце (рис. 7.7, диапазон A2:B9). Далее откроем меню инструмента «*Анализ данных*». Выбираем инструмент «*Регрессия*». Заполним диалоговое окно, как показано на рис. 7.7.



	A	B
1	**x**	**y**
2	0	6
3	1	7,2
4	2	9,4
5	4	11
6	6	15,2
7	8	16,6
8	9	19,4
9	10	21,2

Рис. 7.7 – Исходные данные задачи и диалоговое окно инструмента «Регрессия»

После нажатия OK, программа отобразит расчеты (рис. 7.8).

Используя оценки $\widetilde{a}_0 = 5{,}8676$ и $\widetilde{a}_1 = 1{,}4765$ (ячейки E17 и E18) параметров регрессии a_0 и a_1, запишем выборочное уравнение парной линейной регрессии: $\bar{y}_x = 5{,}8676 + 1{,}4765x$.

	D	E	F	G	H	I	J
1	ВЫВОД ИТОГОВ						
2							
3	*Регрессионная статистика*						
4	Множественный R	0,9939					
5	R-квадрат	0,9878					
6	Нормированный R-квадрат	0,9858					
7	Стандартная ошибка	0,6762					
8	Наблюдения	8					
9							
10	Дисперсионный анализ						
11		*df*	*SS*	*MS*	*F*	*Значимость F*	
12	Регрессия	1	222,3565	222,3565	486,2856	5,68391E-07	
13	Остаток	6	2,7435	0,4573			
14	Итого	7	225,1				
15							
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
17	Y-пересечение	5,8676	0,4114	14,2635	7,42674E-06	4,8610	6,8742
18	x	1,4765	0,0670	22,0519	5,68391E-07	1,3126	1,6403

Рис. 7.8 Результаты анализа линейной модели регрессии

Выборочный коэффициент корреляции $r_{\text{В}} = 0{,}9939 > 0{,}7$, следовательно, связь между изучаемыми признаками в данной совокупности тесная. Коэффициент детерминации $R^{*2} = 0{,}9858$ показывает, что расчетные параметры модели на 98,58% объясняют зависимость между изучаемыми параметрами. Близкий к единице коэффициент детерминации, очень большое расчетное значение $F_{\text{набл}} = 486{,}2856$ статистики F и ничтожно малая статистическая значимость $p \equiv \text{Значимость}\ F = 5{,}68391 \cdot 10^{7}$ свидетельствуют о *высокой адекватности* линейной модели.

Анализ верхней и нижней границ доверительных интервалов (ячейки I17:J17 и I18:J18) приводит к выводу о том, что с вероятностью $\gamma = 1 - \alpha = 0{,}95$ параметры a_0 и a_1, находясь в указанных границах, не принимают нулевых значений, т.е. не являются статистически значимыми и существенно отличны от нуля.

Оценим с помощью средней ошибки аппроксимации качество уравнения. Воспользуемся результатами регрессионного анализа, представленного на рис. 7.9.

ВЫВОД ОСТАТКА

Наблюдение	*Предсказанное y*	*Остатки*
1	5,867647059	0,132352941
2	7,344117647	-0,144117647
3	8,820588235	0,579411765
4	11,77352941	-0,773529412
5	14,72647059	0,473529412
6	17,67941176	-1,079411765
7	19,15588235	0,244117647
8	20,63235294	0,567647059

Рис. 7.9 – Результат применения инструмента «Регрессия» (Вывод остатка)

Составим новую таблицу, как показано на рис. 7.10. В столбце E рассчитаем относительную ошибку аппроксимации по формуле:

$$A_i = \left| \frac{y - \bar{y}_x}{y} \right| \cdot 100\% .$$

Рассчитаем среднюю ошибку аппроксимации: $\overline{A} = \frac{1}{n} \sum A_i = \frac{30{,}95728}{8} \approx 3{,}9$.

Качество построенной модели оценивается как хорошее, так как $\overline{A}$ не превышает 8%.

E2 =ABS(D2/B2)*100

	A	B	C	D	E
1	**Наблюдение**	**y**	**Предсказанное Y**	**Остатки**	**A**
2	1	6	5,86765	0,13235	2,20588
3	2	7,2	7,34412	-0,14412	2,00163
4	3	9,4	8,82059	0,57941	6,16395
5	4	11	11,77353	-0,77353	7,03209
6	5	15,2	14,72647	0,47353	3,11533
7	6	16,6	17,67941	-1,07941	6,50248
8	7	19,4	19,15588	0,24412	1,25834
9	8	21,2	20,63235	0,56765	2,67758
10	**Итого**				**30,95728**
11	**Среднее значение**				**3,86966**

Рис. 7.11 Расчет средней ошибки аппроксимации

В пакете *Excel* для построения аппроксимирующих функций или регрессий можно применить добавление выбранных регрессий (*линий тренда* – trendlines) на диаграмму, построенную на основе таблицы экспериментальных данных исследуемого процесса. Для этого введем исходные данные на новый лист Excel. По этим данным построим точечную диаграмму.

Затем щелкнем правой кнопкой мыши по ряду данных и в появившемся контекстном меню выберем команду «*Добавить линию тренда*» (или после построения на основе ряда данных диаграмму; в меню *Макет* выбрать *Линия тренда*). На экране появится окно *Линия тренда*.

Выберем тип линии тренда – *Линейная* и установим флажки:

- *показывать уравнение на диаграмме*;
- *поместить на диаграмму величину достоверности аппроксимации (R^2)*.

После нажатия кнопки «*Закрыть*» на графике будет показана линия тренда и ее уравнение. Уравнение и коэффициент детерминации можно выделить щелчком левой кнопки мыши и перетащить на то место графика, где их лучше видно (рис. 7.12).

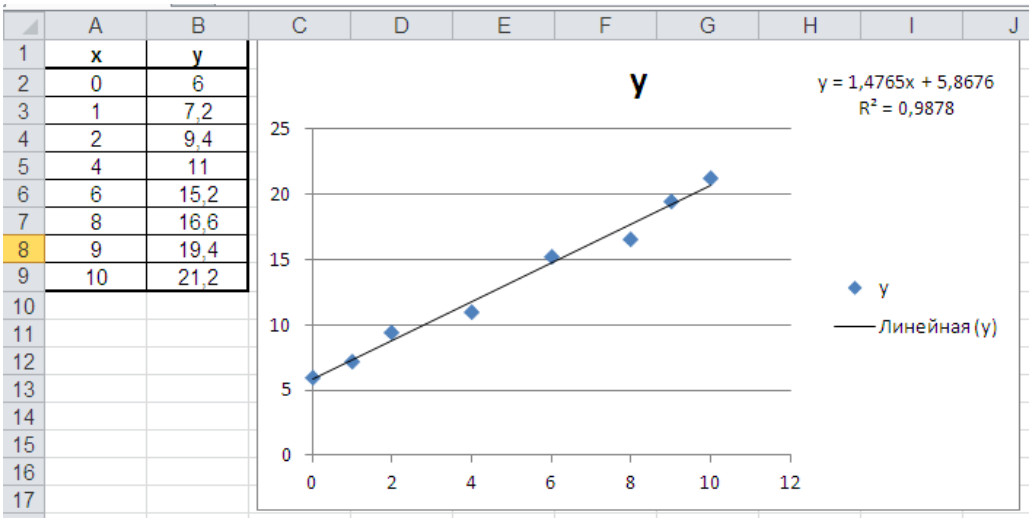


Рис. 7.12 – Вид рабочего листа Excel. Линейный тренд.

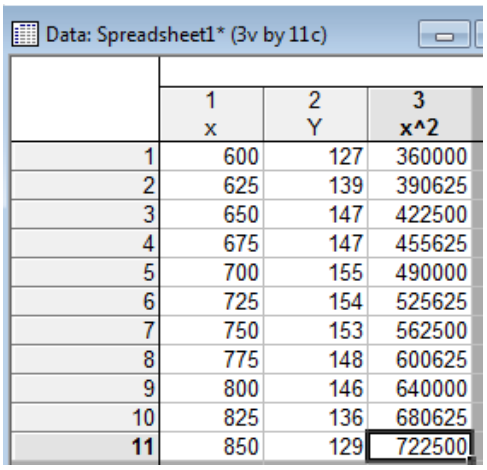
Контрольный пример 7.2. В табл. 7.1 приведены данные, полученные в результате эксперимента, целью которого являлось определение тесноты связи между объемом выпуска продукции (Y) и температурой определенного технологического процесса (x) (см. лабораторную работу 6, контрольный пример 6.3):

Таблица 7.1

x	600	625	650	675	700	725	750	775	800	825	850
Y	127	139	147	137	155	154	153	148	146	136	129

Аппроксимировать статистическую связь между Y и x многочленом второго порядка. Проверить значимость модели. Определить коэффициент детерминации и остаточную дисперсию, сделать вывод. Найти доверительные интервалы для параметров модели. Задание выполнить в пакетах *Excel* и *Statistica*.

*Решение***.** Введём исходные данные в пакете *Statistica* (рис. 7.13):



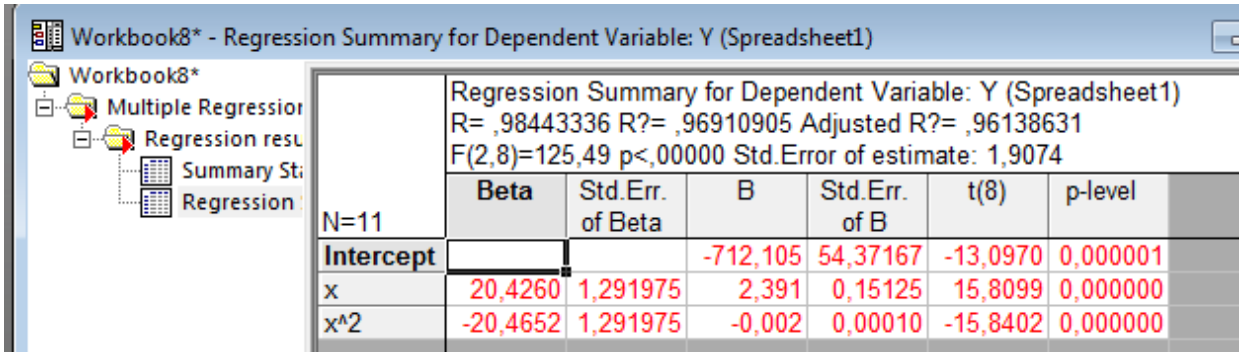
	1 x	2 Y	3 x^2
1	600	127	360000
2	625	139	390625
3	650	147	422500
4	675	147	455625
5	700	155	490000
6	725	154	525625
7	750	153	562500
8	775	148	600625
9	800	146	640000
10	825	136	680625
11	850	129	722500

Рис. 7.13 – Исходные данные

Работаем в модуле «Множественная регрессия»:

Statistics–Multiple Regression---Variables; Dependent var: *Y – Independent var: x, x^2 –Ok – Input File – Raw Data – OK*. В окне *Multiple Regression Results* имеем основные результаты. Нажав кнопку *Summary: Regression result*, получим таблицу результатов (рис. 7.14).

В столбцах приведены: B – значения оценок неизвестных коэффициентов регрессии; *St.Err. of B* – стандартные ошибки оценки коэффициентов; t – значение статистики Стьюдента для проверки гипотезы о нулевом значении коэффициента; *p-level* – уровень значимости отклонения этой гипотезы.



Regression Summary for Dependent Variable: Y (Spreadsheet1)
R= ,98443336 R?= ,96910905 Adjusted R?= ,96138631
F(2,8)=125,49 p<,00000 Std.Error of estimate: 1,9074

N=11	Beta	Std.Err. of Beta	B	Std.Err. of B	t(8)	p-level
Intercept			-712,105	54,37167	-13,0970	0,000001
x	20,4260	1,291975	2,391	0,15125	15,8099	0,000000
x^2	-20,4652	1,291975	-0,002	0,00010	-15,8402	0,000000

Рис. 7.14 – Результаты регрессионного анализа

Аппроксимирующая кривая $\bar{y}_x = -712{,}05 + 2{,}391x - 0{,}002x^2$.

Значение коэффициента детерминации $R^{*2} = 0{,}961386$ (рассматривается именно скорректированный коэффициент детерминации, так как $n < 30$) достаточно велико. Следовательно, полученная эмпирическая функция достаточно точно описывает зависимость Y от x.

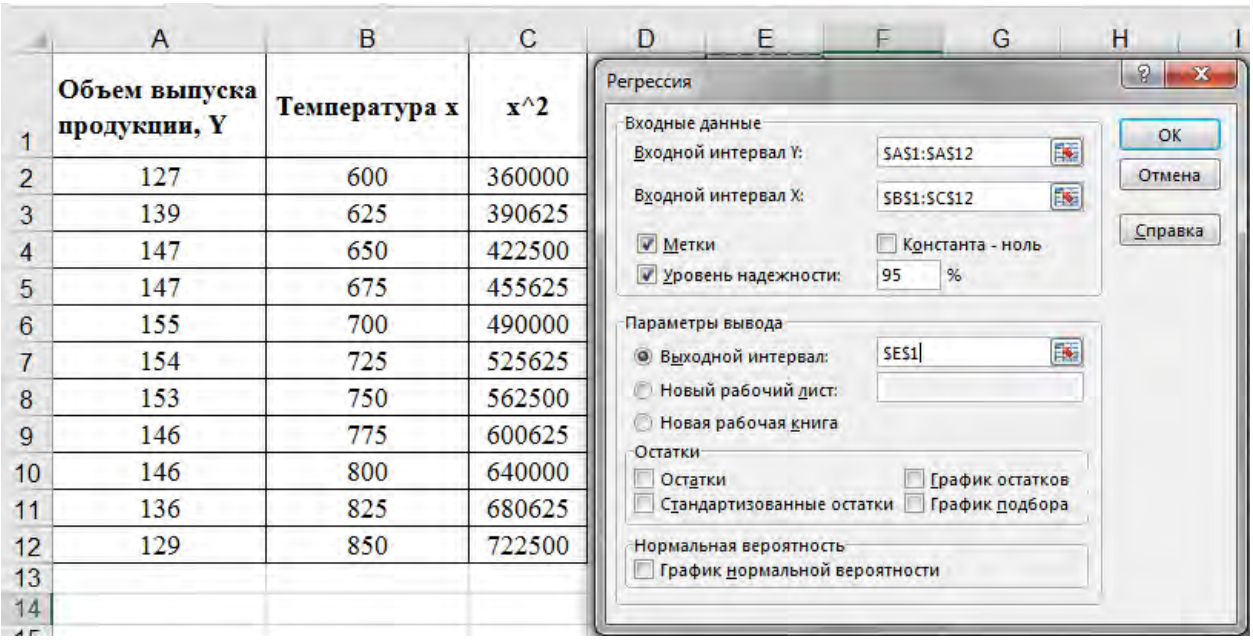
Так как $\sigma_{\text{ост}} = 1{,}9074$ (*Standard error of estimate* – стандартная ошибка оценки), то $D_{\text{ост}} = 1{,}9074^2 \approx 3{,}638$.

Работаем в пакете *Excel*.

Проведём анализ полиномиальной регрессии $f(x) = a_0 + a_1x + a_2x^2$ с .помощью статистической процедуры «*Регрессия*».

Введем исходные данные (рис. 7.15).

Воспользуемся командой *Данные – Анализ данных*. В открывшемся окне выделим процедуру *Регрессия* и щёлкнем на кнопке OK. Заполним диалоговое окно процедуры, как показано на рисунке 7.15.



	A	B	C
1	**Объем выпуска продукции, Y**	**Температура x**	**x^2**
2	127	600	360000
3	139	625	390625
4	147	650	422500
5	147	675	455625
6	155	700	490000
7	154	725	525625
8	153	750	562500
9	146	775	600625
10	146	800	640000
11	136	825	680625
12	129	850	722500
13			
14			

Рис. 7.15 – Диалоговое окно процедуры *Регрессия*

Щелчком на кнопке OK запустим процедуру *Регрессия*. На данном рабочем листе появятся три таблицы результатов реализации процедуры (рис. 7.16).

Использовав оценки $\tilde{a}_0 = -712{,}105$; $\tilde{a}_1 = 2{,}391$; $\tilde{a}_2 = -0{,}0017$ (ячейки F17, F18, F19) параметров регрессии a_0, a_1, a_2, запишем выборочное уравнение полиномиальной регрессии $\bar{y}_x = -712{,}105 + 2{,}391x - 0{,}0017x^2$.

Близкий к единице коэффициент детерминации (ячейка F6), большое расчётное значение статистики F (ячейка I12) и малая значимость F (ячейка J12) свидетельствуют о высокой адекватности полиномиальной модели.

E	F	G	H	I	J	K
ВЫВОД ИТОГОВ						
Регрессионная статистика						
Множественный R	0,9784					
R-квадрат	0,9572					
Нормированный R-квадра	0,9465					
Стандартная ошибка	2,2285					
Наблюдения	11					
Дисперсионный анализ						
	df	*SS*	*MS*	*F*	*Значимость F*	
Регрессия	2	888,999	444,499	89,50700556	3,3486E-06	
Остаток	8	39,729	4,966			
Итого	10	928,727				
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	-699,6098	63,5246	-11,0132	4,1114E-06	-846,0977	-553,1219
Температура x	2,3573	0,1767	13,3401	9,5318E-07	1,9498	2,7648
x^2	-0,0016	0,0001	-13,3741	9,3471E-07	-0,0019	-0,0013

Рис. 7.16 – Результаты анализа полиномиальной модели регрессии

Большие расчётные значения статистики T (ячейки H17, H18, H19) и крайне малые значения p – значимости (ячейки I17, I18, I19) свидетельствуют о том, что выборочные коэффициенты регрессии a_0, a_1, a_2 значимо отличаются от нуля. Об этом же свидетельствуют и доверительные интервалы для коэффициентов регрессии (ячейки I17, J17, I18, J18, I19, J19) соответствующие доверительной вероятности $\gamma = 0{,}95$. Так как доверительные интервалы для коэффициентов a_0, a_1, a_2 не содержат нулевое значение, то эти коэффициенты значимо (существенно) отличаются от нуля.

Полиномиальная регрессия полезна для описания характеристик, имеющих несколько ярко выраженных экстремумов (максимумов и минимумов). Выбор степени полинома определяется количеством экстремумов исследуемой характеристики.

Контрольный пример 7.3. В таблице находится выборка (x, y).

X	1	2	3	4	5	6
Y	10	13.4	15.4	16.5	18.6	19.1

Необходимо:

1) Построить корреляционное поле;

2) По виду полученной диаграммы подобрать 3-4 типа функциональных зависимостей. Рекомендуется выбирать функции с двумя линейно входящими параметрами, например $y = a_0 + a_1 \ln x$, $y = a_0 + \dfrac{a_1}{x}$, $y = a_1\sqrt{x} + a_0$, $y = a_0 + a_1 e^x$;

3) Провести аппроксимацию методом наименьших квадратов;

4) Оценить результаты аппроксимации. Для каждого из полученных эмпирических формул вычислить сумму (невязку) $\varepsilon = \sum_{i}(Y_i - y(x_i))^2$. Сравнивая эти суммы, выбрать эмпирическую формулу, которая более точно описывает результаты эксперимента. Вычислить коэффициент детерминации и остаточную дисперсию.

Задания выполнить в пакете *Mathcad*.

Решение.

1) Введём исходные данные и построим точечный график зависимости (*корреляционное поле*) (рис. 7.17):

$x_0 := 1$ $x_1 := 2$ $x_2 := 3$ $x_3 := 4$ $x_4 := 5$ $x_5 := 6$

$y_0 := 10$ $y_1 := 13.4$ $y_2 := 15.4$ $y_3 := 16.5$ $y_4 := 18.6$ $y_5 := 19.1$

Длина исходных массивов $n := length(x) = 6$ $i := 0..n - 1$

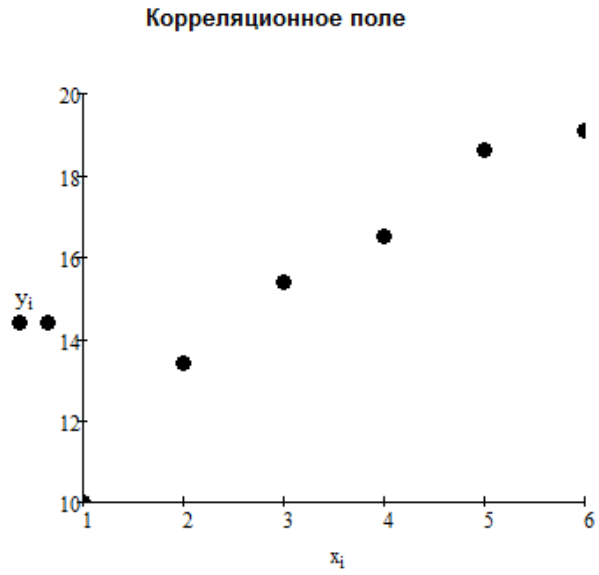


Рис.7.17. Построение корреляционного поля

Из графика видно, что в качестве аппроксимирующей функции можно выбрать нелинейную модель.

Замечание 7.1. Для того чтобы в пакете *Mathcad* получить точечную диаграмму, необходимо щелкнуть по графику два раза мышкой, в появившемся диалоговом окне перейти на вкладку «Трассировка» и выполнить следующие действия (см. рис. 7.18)

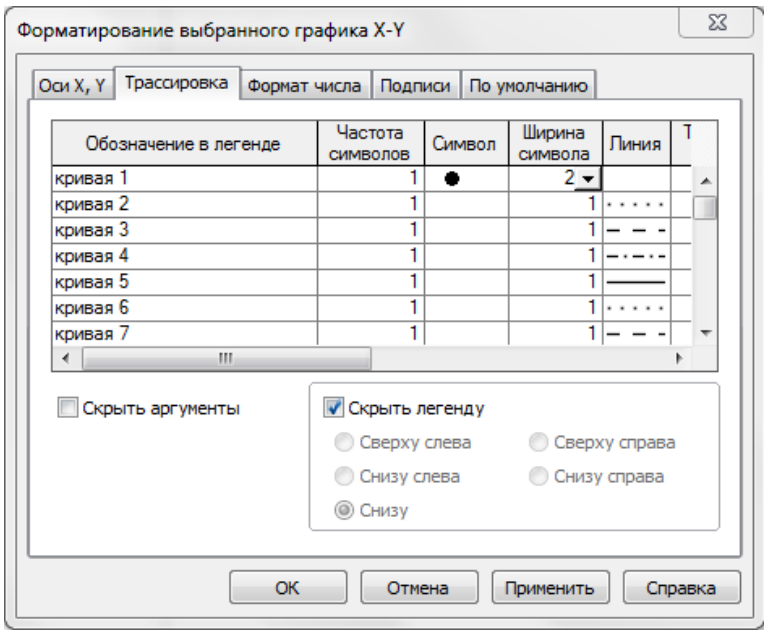


Рис. 7.18 – Форматирование графика

2) Рассмотрим гиперболическую зависимость $\varphi(x) = a_0 + \frac{a_1}{x}$.

После линеаризации найдем параметры этой зависимости, используя формулы (4.9) из теоретического раздела (рис. 7.19).

Замена (линеаризация исходной зависимости)

$$t_i := \frac{1}{x_i} \qquad t = \begin{pmatrix} 1 \\ 0.5 \\ 0.333 \\ 0.25 \\ 0.2 \\ 0.167 \end{pmatrix}$$

$$tcp := \frac{1}{n} \cdot \left(\sum_i t_i \right) = 0.408 \qquad ycp := \frac{1}{n} \cdot \left(\sum_i y_i \right) = 15.5$$

$$Kty := \frac{1}{n-1} \cdot \left[\sum_i \left[\left(t_i - tcp \right) \cdot \left(y_i - ycp \right) \right] \right] = -1.023$$

$$S2 := \frac{1}{n-1} \cdot \left[\sum_i \left(t_i - tcp \right)^2 \right] = 0.098$$

Вычисление параметров a0 и a1

$$a_1 := \frac{Kty}{S2} = -10.415 \qquad a_0 := ycp - a_1 \cdot tcp = 19.753$$

Рис. 7.19 – Вычисление параметров для гиперболической зависимости

Вычислим невязку для данной зависимости:

Вычисление невязки исходной зависимости

Исходная зависимость имеет вид $\varphi 1(x) := a_0 + \frac{a_1}{x}$

$$\varepsilon 1 := \sum_i \left(y_i - \varphi 1\left(x_i\right)\right)^2 = 4.986$$

Остаточная дисперсия и коэффициент детерминации:

Остаточная дисперсия $Doct := \frac{\varepsilon 1}{n - 2} = 1.247$

Коэффициент детерминации $R2 := 1 - \frac{Doct}{\frac{1}{n-1}\cdot\left[\sum_i \left(y_i - ycp\right)^2\right]} = 0.893$

Аналогично получаем:

3) Логарифмическая аппроксимация $\varphi(x) = a_0 + a_1 \ln x$.

Линеаризация:	Получение параметров a_0 и a_1
$t_i := \ln\left(x_i\right)$ $\quad t = \begin{pmatrix} 0 \\ 0.693 \\ 1.099 \\ 1.386 \\ 1.609 \\ 1.792 \end{pmatrix}$	$tcp := \frac{1}{n}\cdot\left(\sum_i t_i\right) = 1.097$ $\quad ycp := \frac{1}{n}\cdot\left(\sum_i y_i\right) = 15.5$ $Kty := \frac{1}{n-1}\cdot\left[\sum_i \left[\left(t_i - tcp\right)\cdot\left(y_i - ycp\right)\right]\right] = 2.252$ $S2 := \frac{1}{n-1}\cdot\left[\sum_i \left(t_i - tcp\right)^2\right] = 0.439$ $a_1 := \frac{Kty}{S2} = 5.129$ $\quad a_0 := ycp - a_1\cdot tcp = 9.876$

Невязка и коэффициент детерминации:

Невязка	Коэффициент детерминации
$\varphi 2(x) := a_0 + a_1 \cdot \ln(x)$ $\varepsilon 2 := \sum_i \left(y_i - \varphi 2\left(x_i\right)\right)^2 = 0.486$	$Doct := \frac{\varepsilon 2}{n-2} = 0.122$ $R2 := 1 - \frac{Doct}{\frac{1}{n-1} \cdot \left[\sum_i \left(y_i - ycp\right)^2\right]} = 0.99$

4) Параболическая зависимость $\varphi(x) = a_0 + a\sqrt{x}_1$.

Линеаризация:	Получение параметров a_0 и a_1
$t_i := \sqrt{x_i}$ $\quad t = \begin{pmatrix} 1 \\ 1.414 \\ 1.732 \\ 2 \\ 2.236 \\ 2.449 \end{pmatrix}$	$tcp := \frac{1}{n} \cdot \left(\sum_i t_i\right) = 1.805$ $\quad ycp := \frac{1}{n} \cdot \left(\sum_i y_i\right) = 15.5$ $Kty := \frac{1}{n-1} \cdot \left[\sum_i \left[\left(t_i - tcp\right) \cdot \left(y_i - ycp\right)\right]\right] = 1.821$ $S2 := \frac{1}{n-1} \cdot \left[\sum_i \left(t_i - tcp\right)^2\right] = 0.289$ $a_1 := \frac{Kty}{S2} = 6.301$ $\quad a_0 := ycp - a_1 \cdot tcp = 4.124$

Невязка и коэффициент детерминации:

Невязка	Коэффициент детерминации
$\varphi 3(x) := a_0 + a_1 \cdot \sqrt{x}$ $\varepsilon 3 := \sum_i \left(y_i - \varphi 3\left(x_i\right)\right)^2 = 0.856$	$Doct := \frac{\varepsilon 3}{n-2} = 0.214$ $R2 := 1 - \frac{Doct}{\frac{1}{n-1} \cdot \left[\sum_i \left(y_i - ycp\right)^2\right]} = 0.982$

5) Экспоненциальная зависимость $\varphi(x) = a_0 + a_1 e^x$.

Линеаризация:	Получение параметров a_0 и a_1
$t_i := \exp(x_i)$ $\quad t = \begin{pmatrix} 2.718 \\ 7.389 \\ 20.086 \\ 54.598 \\ 148.413 \\ 403.429 \end{pmatrix}$	$tcp := \frac{1}{n}\cdot\left(\sum_i t_i\right) = 106.105$ $\quad ycp := \frac{1}{n}\cdot\left(\sum_i y_i\right) = 15.5$ $Kty := \frac{1}{n-1}\cdot\left[\sum_i \left[(t_i - tcp)\cdot(y_i - ycp)\right]\right] = 386.909$ $S2 := \frac{1}{n-1}\cdot\left[\sum_i (t_i - tcp)^2\right] = 2.414 \times 10^4$ $a_1 := \frac{Kty}{S2} = 0.016$ $\quad a_0 := ycp - a_1\cdot tcp = 13.799$

Невязка и коэффициент детерминации:

Невязка	Коэффициент детерминации
$\varphi4(x) := a_0 + a_1\cdot\exp(x)$ $\varepsilon4 := \sum_i \left(y_i - \varphi4(x_i)\right)^2 = 27.228$	$Doct := \frac{\varepsilon4}{n-2} = 6.807$ $R2 := 1 - \frac{Doct}{\frac{1}{n-1}\cdot\left[\sum_i (y_i - ycp)^2\right]} = 0.416$

Внесём основные результаты в следующую таблицу:

№	**Вид зависимости**	**Невязка**	**Коэффициент детерминации**
1	$\varphi_1(x) = 19{,}753 - \frac{10{,}415}{x}$	4,986	0,893
2	$\varphi_2(x) = 9{,}876 + 5{,}129\ln x$	0,486	0,9896
3	$\varphi_3(x) = 4{,}124 + 6{,}301\cdot\sqrt{x}$	0,856	0,9816
4	$\varphi_4(x) = 13{,}779 + 0{,}016\cdot e^x$	27,228	0,4156

Анализируя полученные результаты, делаем вывод, что наиболее точно описывает зависимость между величинами X и Y логарифмическая функция (наименьшая невязка и наибольший коэффициент детерминации).

Задания для самостоятельной работы

Задание 1. Используя данные своего варианта из лабораторной работы № 6 (задание 1):

1) Аппроксимировать статистическую зависимость между этими величинами линейной функцией $\bar{y}_x = a_1 x + a_0$ проверить модель на значимость (адекватность).
2) Вычислить коэффициент детерминации, сделать вывод.
3) Построить корреляционное поле и линию регрессии на корреляционном поле. В пакете *Statistica* провести анализ остатков.

Принять $\alpha = 0{,}05$.

Задания выполнить в пакетах *Statistica* и *Excel*.

Задание 2. Используя данные своего варианта из лабораторной работы № 5 (задание 3) найти оценки параметров модели $\bar{y}_x = a_0 + a_1 x + a_2 x^2$.

Проверить значимость модели. Определить коэффициент детерминации. Найти доверительные интервалы для параметров модели. Принять $\alpha = 0{,}05$.

Построить корреляционное поле и линию регрессии на корреляционном поле. Задание выполнить в пакетах *Excel* и *Statistica*.

Задание 3. В таблице находится выборка (*x*, *y*). Необходимо:

1) Построить корреляционное поле;

2) По виду полученной диаграммы подобрать 3-4 типа функциональных зависимостей. Рекомендуется выбирать функции с двумя линейно входящими параметрами, например $y = a_0 + a_1 \ln x$, $y = a_0 + \frac{a_1}{x}$, $y = a_1\sqrt{x} + a_0$, $y = a_0 + a_1 e^x$;

3) Провести аппроксимацию методом наименьших квадратов;

4) Оценить результаты аппроксимации. Для каждого из полученных эмпирических формул вычислить сумму $S = \sum_i \left(Y_i - y(x_i)\right)^2$. Сравнивая эти суммы, выбрать эмпирическую формулу, которая более точно описывает результаты эксперимента. Вычислить коэффициент детерминации.

Задания выполнить в пакете *Mathcad.*

1	*X*	1	2	3	4	5	6	7	8	9	10
	Y	16,5	13,75	13,31	12,5	13,52	12,75	12,3	12,83	12,28	12,34
2	*X*	0	10	11	16	21	27	32	37	43	48
	Y	8,4	6,2	5,6	5,1	4,2	3,4	3,1	2,5	2,1	1,9
3	*X*	1	2	3	4	5	6	7	8	9	10
	Y	2,11	2,45	2,61	2,73	2,75	2,81	2,87	2,91	2,96	3,03
4	*X*	0,9	1,2	1,5	1,7	2	2,3	2,5	2,8	3,1	3,4
	Y	8,16	3,39	2,19	1,34	0,88	0,61	0,54	0,33	0,28	0,19
5	*X*	0,4	0,8	1,3	1,8	2,2	2,7	3,1	3,6	4,1	4,5
	Y	17	8,8	6,6	5,6	5	4,6	4,3	4,1	3,9	3,8
6	*X*	1	1,5	2	2,5	3	3,5	4	4,5	5	5,5
	Y	4,11	4,16	4,23	4,29	4,36	4,42	4,53	4,58	4,65	4,73

7	*X*	3	4	5	6	7	8	9	10	11	12
	Y	0,43	0,51	0,62	0,81	1,01	1,23	1,47	1,53	1,75	2,25
8	*X*	0,3	1,57	2,84	4,11	5,38	6,65	7,92	9,19	10,46	11,73
	Y	15,33	4,55	3,41	2,97	2,74	2,6	2,59	2,44	2,38	2,34
9	*X*	0,01	0,04	0,07	0,1	0,13	0,16	0,19	0,22	0,25	0,28
	Y	-2,04	-1,35	-1,07	-0,89	-0,76	-0,66	-0,57	-0,5	-0,44	-0,38
10	*X*	0,5	1	1,5	2	2,5	3	3,5	4	4,5	5
	Y	0,24	-0,02	-0,11	-0,15	-0,18	-0,19	-0,2	-0,21	-0,22	-0,23
11	*X*	0,15	0,94	1,72	2,51	3,29	4,08	4,86	5,65	6,43	7,22
	Y	-9,69	-4,2	-2,37	-1,25	-0,43	0,21	0,74	1,3	1,58	1,93
12	*X*	1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	1,9	2
	Y	0,460	0,613	0,702	0,800	0,908	1,028	1,160	1,307	1,468	1,647

Лабораторная работа № 8. Непараметрические методы математической статистики

Необходимые теоретические сведения для выполнения лабораторной работы находятся в теоретическом разделе – тема 5.1.

Контрольный пример 8.1. Предполагается, что один из двух приборов, определяющих скорость автомобиля, имеет систематическую ошибку. Для проверки этого предположения определили скорость 10 автомобилей, причём скорость каждого фиксировалась одновременно двумя приборами.

В результате получены следующие данные:

v_1, км/ч	70	85	63	54	65	80	75	95	52	55
v_2, км/ч	72	86	62	55	63	80	78	90	53	57

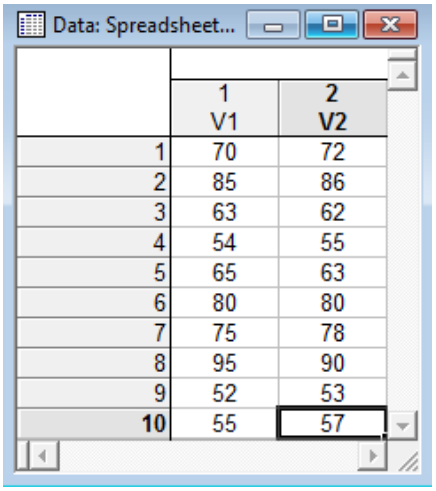
Позволят ли эти результаты утверждать, что второй прибор действительно даёт завышенные значения скорости? Принять $\alpha = 0{,}1$. Задачу решить с помощью пакета *Statistica*, используя критерий знаков и знако-ранговый критерий Вилкоксона.

Решение. Критерий знаков является непараметрической альтернативой *t*-критерию Стьюдента в случае зависимых выборок, который применяется, когда проводится два измерения (например, в различных условиях) одних и тех же объектов и необходимо установить наличие или отсутствие различия результатов.

Для применения этого критерия требуются очень слабые предположения (например, однозначная определенность медианы для разности значений).

При нулевой гипотезе (отсутствие эффекта обработки) число положительных разностей имеет биномиальное распределение со средним, равным половине объема выборки, основываясь на этом можно вычислить критические значения.

Введем исходные данные (рис. 8.1):



	1 V1	2 V2
1	70	72
2	85	86
3	63	62
4	54	55
5	65	63
6	80	80
7	75	78
8	95	90
9	52	53
10	55	57

Рис. 8.1 – Таблица исходных данных

Для запуска модуля *Непараметрические статистики* в меню *Statistics* необходимо выбрать *Nonparametrics*, стартовая панель изображена на рис. 8.2

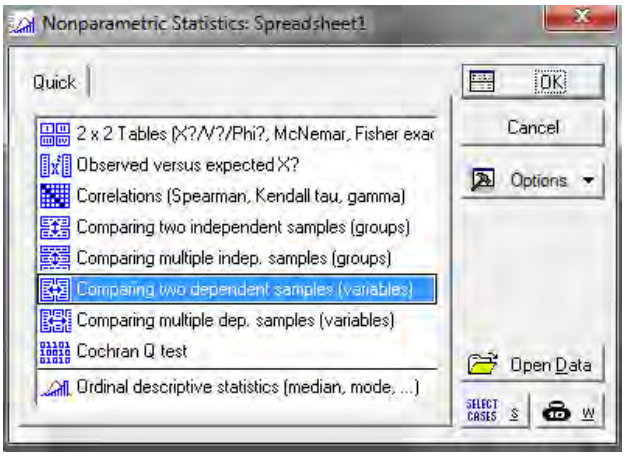


Рис. 8.2 – Стартовая панель модуля *Nonparametrics*

Запустим модуль непараметрических статистик и выберем в нем процедуру *Comparing two dependent samples (variables)*. В открывшемся окне зададим переменные для первого и второго списков (кнопка *Variables*):

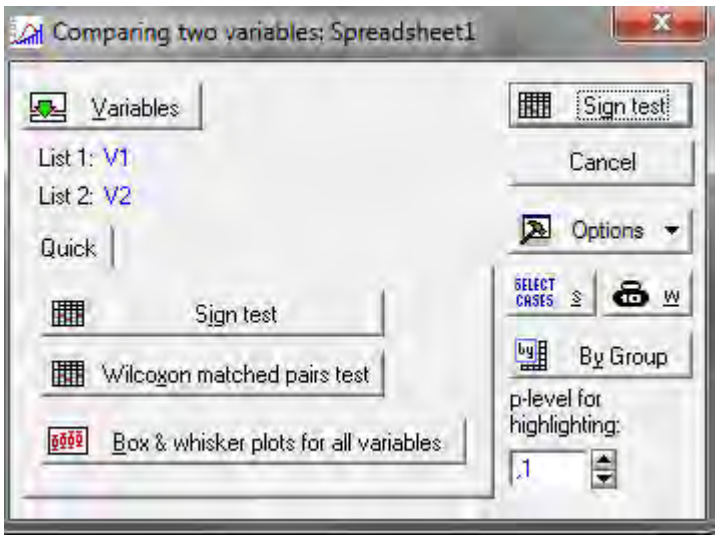


Рис. 8.3 – Задание переменных

Нажав на кнопку *Sign test*, рассчитаем характеристики для критерия знаков:

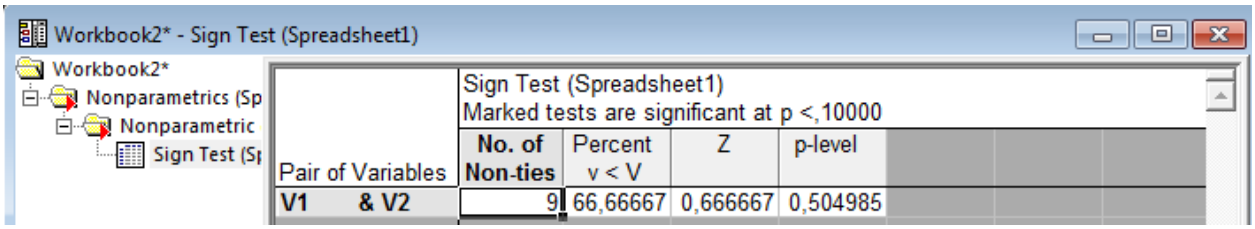
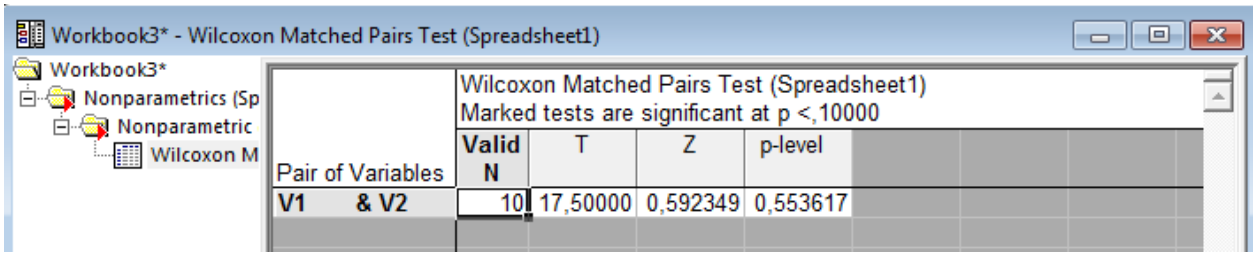


Рис. 8.4 – Результаты критерия знаков

Первый столбец содержит названия сравниваемых групп (в нашем случае V1 и V2), второй – измерение скорости шестого автомобиля обоими приборами игнорируется, т.к. оно дало одинаковый результат), пятый – уровень значимости. Из заголовка таблицы следует, для наличия значимых различий между группами уровень значимости должен быть меньше 0,1 (в случае нашего примера он равняется 0,504985). Это означает, что различие между результатами измерений каждым из приборов не является значимым.

Знако-ранговый критерий Вилкоксона также является непараметрической альтернативой *t*-критерию в случае зависимых выборок. При этом предполагается, что рассматриваемые переменные ранжированы. Требования к критерию Вилкоксона более строгие, чем к критерию знаков. Однако если они удовлетворены, то критерий Вилкоксона имеет большую мощность, чем критерий знаков.

Проверим различия между результатами измерений по критерию Вилкоксона, нажав кнопку *Wilcoxon matches pair test*:



Wilcoxon Matched Pairs Test (Spreadsheet1)
Marked tests are significant at p <,10000

Pair of Variables	Valid N	T	Z	p-level
V1 & V2	10	17,50000	0,592349	0,553617

Рис. 8.5 – Результаты критерия Вилкоксона

Как видно из рис. 8.5, уровень значимости равен 0,553617 и также значительно отличается от 0,1. Таким образом, вывод аналогичен предыдущему.

Проиллюстрируем полученные выводы с помощью диаграммы размаха, нажав соответствующую кнопку в окне *Box & whisker plots for all variables,* представленном на рис.8.3.

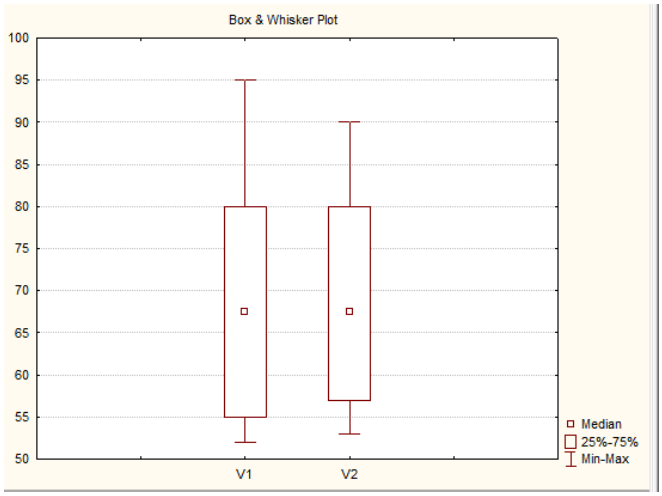


Рис. 8.6 – Диаграмма размаха

На диаграмме размаха для каждой переменной показаны: медиана, квартильный размах (25% и 75%), размах (минимум, максимум).

Контрольный пример 8.2. Проверить на уровне значимости $\alpha = 0{,}05$ гипотезу H_0 об однородности двух выборок (наблюдаемые различия между значениями признака в рассматриваемых выборках случайны), объемы которых $n_1 = 9$, $n_2 = 8$ (в первой строке приведены варианты первой выборки, во второй строке – варианты второй выборки).

x	23	31	27	28	27	39	21	40	35
y	30	49	32	26	52	36	26	50	

Задачу решить с применением пакета *Statistica*, используя критерии Манна-Уитни, Вальда-Вольфовица и Колмогорова-Смирнова.

Решение.

H_0: Наблюдаемые различия между значениями признака в рассматриваемых выборках случайны.

H_1: Наблюдаемые различия между значениями признака в рассматриваемых выборках не случайны.

Введём исходные данные (рис. 8.7):

	1 n	2 x
1	1	23
2	1	31
3	1	27
4	1	28
5	1	27
6	1	39
7	1	21
8	1	40
9	1	35
0	2	30
1	2	49
2	2	32
3	2	26
4	2	52
5	2	36
6	2	26
7	2	50

Рис. 8.7 – Таблица исходных данных.

Запустим модуль непараметрических статистик (*Statistics – Nonparametrics*) и выберем в нем процедуру *Comparing two independent samples (groups)*. Зададим зависимую и группирующую переменные, нажав на кнопку *Variables*. В данном примере зависимой является переменная x, группирующей – n.

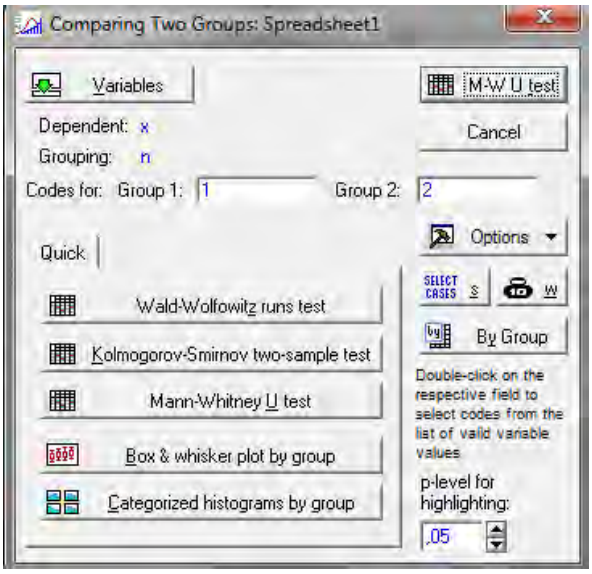
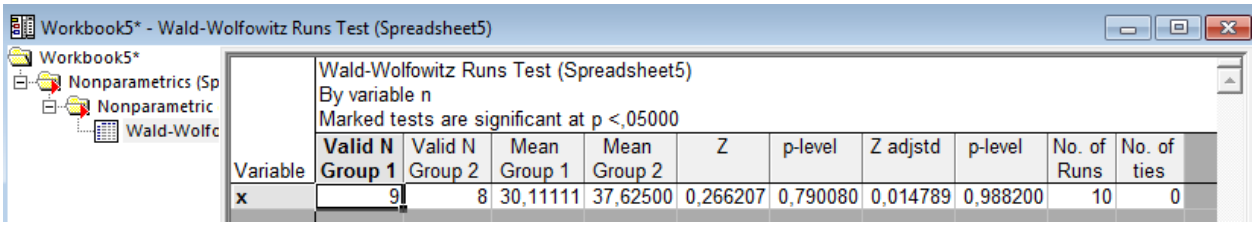


Рис. 8.8 – Задание зависимой и группирующей переменных

На этой же панели в виде кнопок отображены все возможные тесты для анализа данных: критерий Вальда-Вольфовица, Колмогорова-Смирнова и Манна-Уитни. Выполним каждый из них, поочередно выбирая соответствующую кнопку и сравним полученные результаты.



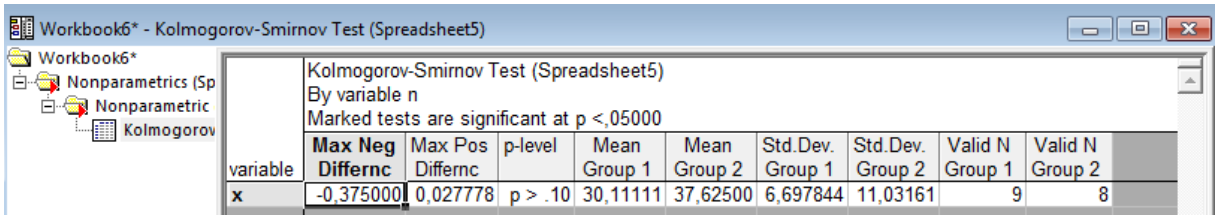
Wald-Wolfowitz Runs Test (Spreadsheet5)
By variable n
Marked tests are significant at p <,05000

Variable	Valid N Group 1	Valid N Group 2	Mean Group 1	Mean Group 2	Z	p-level	Z adjstd	p-level	No. of Runs	No. of ties
x	9	8	30,11111	37,62500	0,266207	0,790080	0,014789	0,988200	10	0

Рис. 8.9 – Результаты теста Вальда-Вольфовица

Первый столбец результирующей таблицы содержит название исследуемого признака, два следующих – количество наблюдаемых измерений по каждому признаку (в данном случае для первой выборки (x) и второй (y)). Два следующих столбца содержат средние значения каждого признака.

Как видно из таблицы результатов, различие между выборками не является значимым: $p = 0{,}9882 > 0{,}05$.



Kolmogorov-Smirnov Test (Spreadsheet5)
By variable n
Marked tests are significant at p <,05000

variable	Max Neg Differnc	Max Pos Differnc	p-level	Mean Group 1	Mean Group 2	Std.Dev. Group 1	Std.Dev. Group 2	Valid N Group 1	Valid N Group 2
x	-0,375000	0,027778	p > .10	30,11111	37,62500	6,697844	11,03161	9	8

Рис. 8.10 – Результаты теста Колмогорова-Смирнова

Здесь: максимальная отрицательная и положительная разности, уровень значимости результатов, средние значения по каждому из признаков, стандартные отклонения для каждого из признаков и количество наблюдаемых измерений по каждому признаку.

Так как $p > 0.1$ (столбец 3), то наблюдаемые различия между значениями признака в рассматриваемых выборках случайны

Можно заметить, что стандартные отклонения в обеих группах не равны (см. рис. 8.9 и рис. 8.10), следовательно, невозможно применить *t*-критерий.

Mann-Whitney U Test (Spreadsheet1)
By variable n
Marked tests are significant at p <,05000

variable	Rank Sum Group 1	Rank Sum Group 2	U	Z	p-level	Z adjusted	p-level	Valid N Group 1	Valid N Group 2	2*1sided exact p
x	68,00000	85,00000	23,00000	-1,25093	0,210963	-1,25246	0,210403	9	8	0,235870

Рис. 8.11 – Результаты критерия Манна-Уитни

Самое главное, на что следует обратить внимание в итоговой таблице теста – величина вероятности ошибки p. При большом числе наблюдений в выборках (20 и более) значение p необходимо искать в 5-м столбце таблицы (вслед за «Z»), иначе – в 7-м (вслед за «Z-adjusted»). При $p < \alpha$ делается вывод о наличии статистически значимой разницы между сравниваемыми выборками.

Так как $p = 0{,}210403 > \alpha = 0{,}05$, то статистически значимой разницы между выборками нет – они однородны, т.е. принадлежат одной генеральной совокупности.

Проиллюстрируем полученные выводы с помощью диаграммы размаха (см. рис. 8.12) – кнопка 

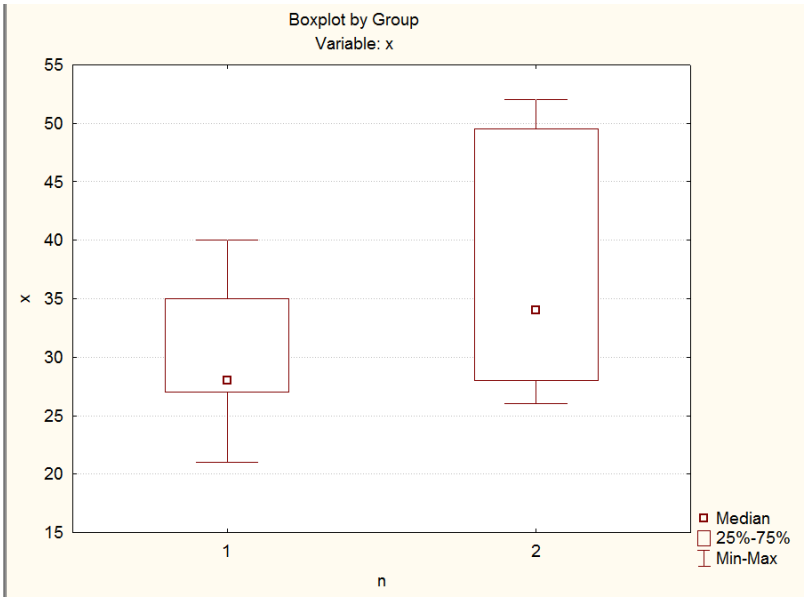


Рис. 8.12 – Диаграмма размаха

Контрольный пример 8.3. Три группы водителей обучались по различным методикам. После окончания срока обучения был произведен тестовый контроль над случайно отобранными водителями из каждой группы. Получены следующие результаты:

№ группы	Число ошибок, допущенных водителями, x_{ij}
1	1 3 2 1 0 2 1
2	2 3 2 1 3 3 1
3	4 2 3 2 1

На уровне значимости $\alpha = 0{,}05$ с помощью критерия Краскелла – Уоллиса проверить гипотезу об отсутствии влияния различных методик обучения на результаты тестового контроля водителей. Задание выполнить в пакете *Statistica*.

Решение. Формулируем нулевую и конкурирующую гипотезу:

- H_0: различные методики обучения не влияют на результаты тестового контроля водителей;
- H_1: различные методики обучения влияют на результаты тестового контроля водителей.

Введём исходные данные (рис. 8.13):

	1 Error	2 Code
1	1	1
2	3	1
3	2	1
4	1	1
5	0	1
6	2	1
7	1	1
8	2	2
9	3	2
10	2	2
11	1	2
12	3	2
13	3	2
14	1	2
15	4	3
16	2	3
17	3	3
18	2	3
19	1	3

Рис. 8.13 – Исходная выборка данных
(Error – ошибки; Code – код)

В стартовой панели модуля *Nonparametrics* выбираем *Comparing multiple indep. samples (groups)*.

В появившемся окне выбираем *Variables* и задаём переменные (рис. 8.14); затем нажимаем OK.

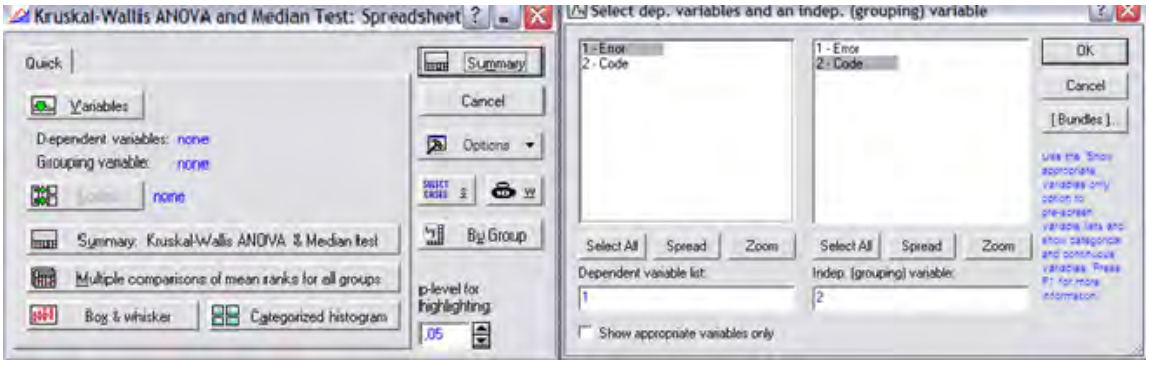


Рис. 8.14 – Окно *Kruskal-Wallis Anova and Median test u* Окно выбора переменных

Далее нажимаем *Codes* и выбираем коды для группируемых переменных, щёлкнув по кнопке *All* (рис. 8.15):

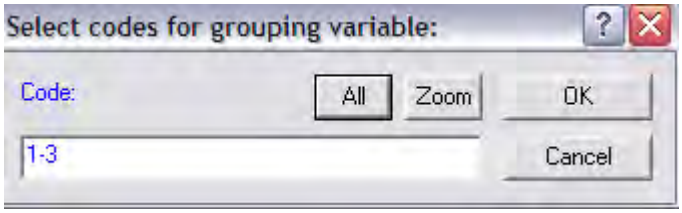
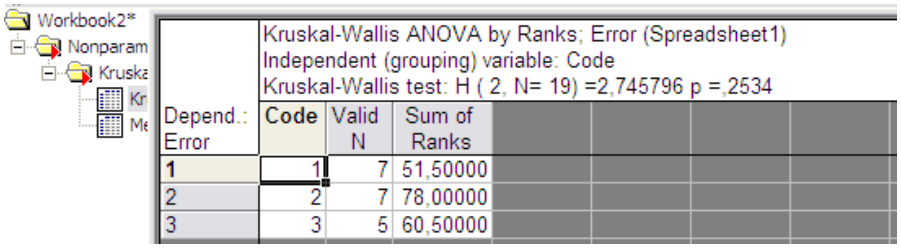


Рис. 8.15 – Окно выбора кода

Нажимаем *Summary* и получаем следующую таблицу результатов (рис. 8.16):



Kruskal-Wallis ANOVA by Ranks; Error (Spreadsheet1)
Independent (grouping) variable: Code
Kruskal-Wallis test: H (2, N= 19) =2,745796 p =,2534

Depend.: Error	Code	Valid N	Sum of Ranks
1	1	7	51,50000
2	2	7	78,00000
3	3	5	60,50000

Рис. 8.16 – Таблица результатов анализа.

Так как p – значение, равное $p = 0{,}2534$ больше уровня значимости $\alpha = 0{,}05$, гипотеза H_0 принимается – разные методики не влияют на результат обучения.

Контрольный пример 8.4. Киноплёнка четырёх видов была представлена трём экспертам для определения лучшей из них. Каждому эксперту предложили упорядочить плёнки по степени предпочтения. Баллы (ранги), поставленные экспертами, приведены в таблице 8.1. Наибольший балл соответствует плёнке самого лучшего качества.

Таблица 8.1

Вид плёнки	**Эксперты**		
	1	**2**	**3**
П1	2	3	2
П2	5	4	5
П3	3	3	3
П4	4	5	5

Требуется, используя критерий Фридмана, определить, различаются ли виды плёнок и согласованы ли оценки экспертов. Задание выполнить в пакете *Statistica*.

Решение. Введём исходные данные (рис. 8.17):

	1 P1	2 P2	3 P3	4 P4
1	2	5	3	4
2	3	4	3	5
3	2	5	3	5

Рис. 8.17 – Исходная выборка данных

В стартовой панели модуля *Nonparametric Statistics* (Непараметрические статистики) выбираем *Comparing multiple dep. samples (variables)*.

В появившемся окне нажимаем *Variables* и задаём переменные, нажав кнопку *Select All* (рис. 8.18):

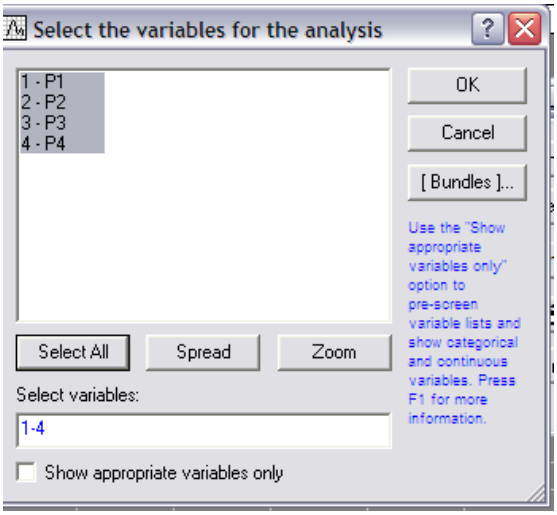


Рис. 8.18 – Окно выбора переменных

В появившемся окне *Friedman ANOVA by ranks* (Двухфакторный анализ Фридмана) нажимаем *Summary* и получаем следующую таблицу результатов (рис. 8.19):

Friedman ANOVA and Kendall Coeff. of Concordance (Spreadsheet1)
ANOVA Chi Sqr. (N = 3, df = 3) = 8,142857 p = ,04315
Coeff. of Concordance = ,90476 Aver. rank r = ,85714

Variable	Average Rank	Sum of Ranks	Mean	Std.Dev.
P1	1,166667	3,50000	2,333333	0,577350
P2	3,500000	10,50000	4,666667	0,577350
P3	1,833333	5,50000	3,000000	
P4	3,500000	10,50000	4,666667	0,577350

Рис. 8.19. – Таблица результатов анализа

Гипотеза H_0 проверяется с помощью статистики Фридмана. Гипотеза отклоняется на уровне значимости α, если $F_{\text{набл}} > \chi^2(\alpha, m-1)$. Значение выборочной статистики в данном случае $F_{\text{набл}} = 8{,}143$, а при $\alpha = 0{,}05$ – $\chi^2(0{,}05;3) = 7{,}815$. Следовательно, гипотеза H_0 отклоняется: следует считать, что виды плёнок, по мнению экспертов, различны.

Мерой согласия различных ранжировок n объектов является *коэффициент конкордации (согласованности) Кендалла* W. В данном случае $W = 0{,}905$. Большое значение W свидетельствует о согласованности оценок экспертов.

Задания для самостоятельной работы

Задание 1. Предполагается, что один из двух приборов, определяющих скорость автомобиля, имеет систематическую ошибку. Для проверки этого предположения определили скорость n автомобилей, причём скорость каждого фиксировалась одновременно двумя приборами.

Позволят ли эти результаты утверждать, что один из приборов действительно даёт завышенные значения скорости? Принять $\alpha = 0{,}05$. Задачу решить с применением пакета *Statistica*. Применить критерий знаков и знако-ранговый критерий.

Вариант 1													
v_1, км/ч	53	83	50	65	58	50	58	79	50	84	70	60	
v_2, км/ч	57	75	55	69	66	72	60	80	51	85	68	79	
Вариант 2													
v_1, км/ч	78	67	58	84	64	81	57	78	65	62	66	56	78
v_2, км/ч	84	85	59	85	63	71	64	65	74	64	76	58	58
Вариант 3													
v_1, км/ч	70	51	66	56	72	75	80	73	80	51	72	65	68
v_2, км/ч	67	59	71	61	69	73	79	79	77	78	68	60	65
Вариант 4													
v_1, км/ч	66	64	78	80	71	80	59	71	62	68	61	60	56
v_2, км/ч	58	67	77	78	67	75	59	74	61	72	63	59	58
Вариант 5													
v_1, км/ч	73	82	67	63	76	68	63	76	71	74	73	82	67
v_2, км/ч	62	85	70	60	76	64	66	82	77	69	72	82	70
Вариант 6													
v_1, км/ч	54	69	53	70	64	69	75	73	65	74	84	69	73
v_2, км/ч	55	78	53	79	54	70	84	71	70	82	85	70	83
Вариант 7													
v_1, км/ч	72	85	74	73	79	59	51	70	53	82	72	85	74
v_2, км/ч	62	79	61	64	88	51	60	70	58	75	72	79	61
Вариант 8													
v_1, км/ч	70	57	73	53	53	58	54	81	60	56	59	57	73
v_2, км/ч	61	53	75	53	52	52	62	73	72	51	51	57	65
Вариант 9													
v_1, км/ч	84	51	54	74	91	62	80	45	77	84	73	79	
v_2, км/ч	81	56	55	88	92	77	73	46	61	82	69	82	

Вариант 10													
v_1, км/ч	59	64	52	82	50	56	48	70	79	61	76	53	49
v_2, км/ч	56	75	57	78	52	56	47	73	80	70	77	53	50
Вариант 11													
v_1, км/ч	64	70	54	55	71	51	53	83	72	59	51	68	54
v_2, км/ч	62	61	53	50	68	51	49	80	75	65	50	57	62
Вариант 12													
v_1, км/ч	75	5	61	50	63	40	58	83	57	76	70	72	
v_2, км/ч	81	84	69	50	66	40	69	90	59	73	61	77	

Задание 2. Используя критерии Вилкоксона и Манна-Уитни, проверить на уровне значимости α гипотезу H_0 об однородности двух выборок (наблюдаемые различия между значениями признака в рассматриваемых выборках случайны). В первой строке приведены варианты первой выборки, во второй строке – варианты второй выборки. Задание выполнить в пакете *Statistica*.

Вариант 1 $\alpha = 0{,}01$													
x_i	2,3	3,3	4,6	2,1	3,4	6,3	1,5	2,7	6,5	4,1	7,1		
y_i	1,3	2,4	4,5	3,2	2,5	4,2	3,5	4,6	2,8				
Вариант 2 $\alpha = 0{,}05$													
x_i	50	80	45	17	81	70	66	55	72	24	34	80	15
y_i	59	79	59	82	60	94	92	54	75	25	57	81	
Вариант 3 $\alpha = 0{,}05$													
x_i	27	21	39	14	11	36	39	23	37	37	27	21	
y_i	12	20	17	16	27	12	21	33	37	15	12	20	
Вариант 4 $\alpha = 0{,}05$													
x_i	38	41	43	46	48	52	56	57	60	65	68	73	
y_i	33	34	36	37	39	40	42	44	45	47	49	51	
Вариант 5 $\alpha = 0{,}05$													
x_i	12	14	15	18	21	25	26	27	30	31	32	35	
y_i	11	13	16	17	19	20	22	23	24	26	28	29	
Вариант 6 $\alpha = 0{,}01$													
x_i	17	10	19	17	33	20	22	8	19	16	20		
y_i	13	6	15	9	12	7	5	17	17	35	25	23	21
Вариант 7 $\alpha = 0{,}05$													
x_i	24	26	27	27	30	32	33	34	35	36			
y_i	21	21	22	23	25	25	25	25	27	27	29	31	

Вариант 8 $\alpha = 0{,}05$													
x_i	11	10	19	18	33	20	22	8	13	6	15	9	
y_i	12	7	5	14	19	17	16	2	35	25	23	21	
Вариант 9 $\alpha = 0{,}05$													
x_i	135	222	251	260	269	235	386	252	352	173	156		
y_i	294	311	286	364	277	336	208	346	239	172	254		
Вариант 10 $\alpha = 0{,}1$													
x_i	28	33	39	40	41	42	45	46	47				
y_i	34	40	41	42	43	44	46	48	49	42			
Вариант 11 $\alpha = 0{,}05$													
x_i	0,09	0,19	0,27	0,35	0,5	0,58	0,62	0,74	0,8	0,91			
y_i	0,12	0,18	0,26	0,37	0,46	0,6	0,66	0,73	0,87	0,94			
Вариант 12 $\alpha = 0{,}01$													
x_i	8	3	5	3	9	5	4	11					
y_i	5	8	5	9	8	4	8	7	12	8			

Задание 3. *m* групп испытуемых выполняли тест в разных экспериментальных условиях. Задача в том, чтобы установить (с помощью критерия Краскелла-Уоллиса) – зависит ли эффективность выполнения теста от условий, или, иными словами, существуют ли статистически достоверные различия между группами.

В таблице приведено число ошибок показателя переключаемости внимания (в процентах).

Задание выполнить в пакете *Statistica*.

Вариант 1					**Вариант 2**			
№	**1 группа**	**2 группа**	**3 группа**	**4 группа**	**№**	**1 группа**	**2 группа**	**3 группа**
1	14	13	10	12	**1**	21	26	18
2	16	10	17	12	**2**	26	27	24
3	10	17	16	17	**3**	20	20	26
4	14	16	14	10	**4**	25	23	20

Вариант 3				**Вариант 4**				
№	**1 группа**	**2 группа**	**3 группа**	**№**	**1 группа**	**2 группа**	**3 группа**	**4 группа**
1	30	26	20	**1**	8	10	11	12
2	20	20	21	**2**	12	9	13	13
3	26	21	19	**3**	9	10	12	
4	20	18	18	**4**	14	10		
5	19	18	16					
6	21	17	17					

Вариант 5					Вариант 6			
№	1 группа	2 группа	3 группа	4 группа	№	1 группа	2 группа	3 группа
1	23	46	34	20	**1**	15	13	6
2	20	12	25	30	**2**	21	28	21
3	33	33	22	26	**3**	24	25	16
4	36	22	12	27	**4**	12	20	9
					5		16	
					6		20	

Вариант 7					Вариант 8			
№	1 группа	2 группа	3 группа	4 группа	№	1 группа	2 группа	3 группа
1	11	17	20	21	**1**	2	11	8
2	10	16	22	30	**2**	10	7	12
3	11	18	21	19	**3**	5	8	14
4	13	19	18	18	**4**	8	12	9
5	11	17	19	20	**5**	10	12	16
6	10	16	19	21	**6**	7	12	14
					7	12	9	10

Вариант 9				Вариант 10			
№	1 группа	2 группа	3 группа	№	1 группа	2 группа	3 группа
1	6	8	10	**1**	48	50	47
2	8	9	9	**2**	46	49	45
3	10	7	8	**3**	42	42	46
4	7	10	10	**4**	41	43	30
5	6	9	10	**5**	37	39	32
6	7	8	10	**6**	32	28	41

Вариант 11				Вариант 12			
№	1 группа	2 группа	3 группа	№	1 группа	2 группа	3 группа
1	32	43	50	**1**	6	5	10
2	38	45	56	**2**	8	1	12
3	31	49	58	**3**	7	10	9
4	40	46	54	**4**	5	3	13
5	39	51	52	**5**	6	6	4

Задание 4. Шести студентам было предложено ответить на вопросы теста. Фиксировалось время решения каждого задания. С помощью критерия Фридмана ответить на вопрос: наблюдаются ли статистически значимые различия между временем решения первых трех заданий теста?

К каждому числу прибавить номер варианта.

№ испытуемых	Время решения 1-го задания теста, в сек.	Время решения 2-го задания теста, в сек.	Время решения 3-го задания теста, в сек.
1	8 + V	3 + V	5 + V
2	4 + V	15 + V	12 + V
3	6 + V	23 + V	15 + V
4	3 + V	6 + V	6 + V
5	7 + V	12 + V	3 + V
6	15 + V	24 + V	12 + V

Лабораторная работа № 9. Прогнозирование временных рядов. Множественная линейная регрессия.

Необходимые теоретические сведения для выполнения лабораторной работы находятся в теоретическом разделе – тема 6.1.

Контрольный пример 9.1. Ниже приведены данные о числе сообщений, переданных в течение 42 суток (6 недель) в одной из радиосетей морской связи.

Таблица 9.1

День недели	**Неделя**					
	1	**2**	**3**	**4**	**5**	**6**
Понедельник	1	3	5	4	6	6
Вторник	2	5	7	5	7	11
Среда	8	6	10	8	12	16
Четверг	10	8	6	9	13	15
Пятница	2	5	5	13	8	8
Суббота	1	3	6	5	6	8
Воскресенье	0	2	2	4	5	6

Используя эти данные (с помощью скользящего среднего и экспоненциального сглаживания), выявить основные тенденции процесса радиопередач в рассматриваемой радиосети.

Решение.

В диапазон A1:A42 листа *Excel* ведем значения временного ряда из приведенной таблицы (на рис. 9.1 виден начальный отрезок этого ряда, цветом выделены элементы ряда, приходящиеся на выходные дни).

На вкладке *Данные* выберем *Анализ данных*. В отрывшемся диалоговом окне выделим процедуру *Скользящее среднее* и щелкнем на кнопку ОК. На экране появится диалоговое окно *Скользящее среднее*.

В поле ввода *Входной интервал* этого окна введем ссылку A1:A42 на диапазон ячеек, содержащий элементы исследуемого временного ряда. В поле *Интервал* введем размер окна сглаживания $m = 7$. В поле ввода *Выходной интервал* введем ссылку B1 на верхнюю ячейку столбца результатов сглаживания и установим флажок *Вывод графика*.

Щелкнем на кнопке ОК.

	A	B	C
1	1	#Н/Д	
2	2	#Н/Д	
3	8	#Н/Д	
4	10	#Н/Д	3,43
5	2	#Н/Д	3,71
6	1	#Н/Д	4,14
7	0	3,43	3,86
8	3	3,71	3,57
9	5	4,14	4
10	6	3,86	4,29
11	8	3,57	4,57
12	5	4	4,86
13	3	4,29	5,14
14	2	4,57	5,71
15	5	4,86	5,43
16	7	5,14	5,43
17	10	5,71	5,86
18	6	5,43	5,86
19	5	5,43	5,71
20	6	5,86	5,43
21	2	5,86	5,14
22	4	5,71	5,57
23	5	5,43	6,71
24	8	5,14	6,57

Рис. 9.1 – Решение примера 9.1 (сглаживание с помощью скользящего среднего)

Справа от столбца с исходными данными в диапазоне B1:B42, появятся столбец адаптивных скользящих средних $\widetilde{y}_i$ и график исследуемого временного ряда с наложенным на него графиком адаптивных скользящих средних $\widetilde{y}_i; i = \overline{7,42}$ (см. рис. 9.1).

На графике временного ряда видны явно выраженные периодические колебания числа радиопередач с периодом, равным семи суткам. На графике адаптивных скользящих средних периодические колебания практически не заметны (это обусловлено тем, что размер окна сглаживания равен периоду колебаний). График адаптивных скользящих средних свидетельствует о медленном росте тренда временного ряда числа радиопередач.

Для сравнения простого и адаптивного скользящих средних в диапазоне C4:C39 приведены значения скользящего среднего, вычисленные по канонической формуле. Эти вычисления выполнены по формуле = СРЗНАЧ(A1:A7), введенной в ячейку C4 и скопированной затем в ячейки диапазона C5:C39. График, построенный по этим значениям, приведен на рис. 9.2. На этом же рисунке приведен график адаптивного скользящего среднего.

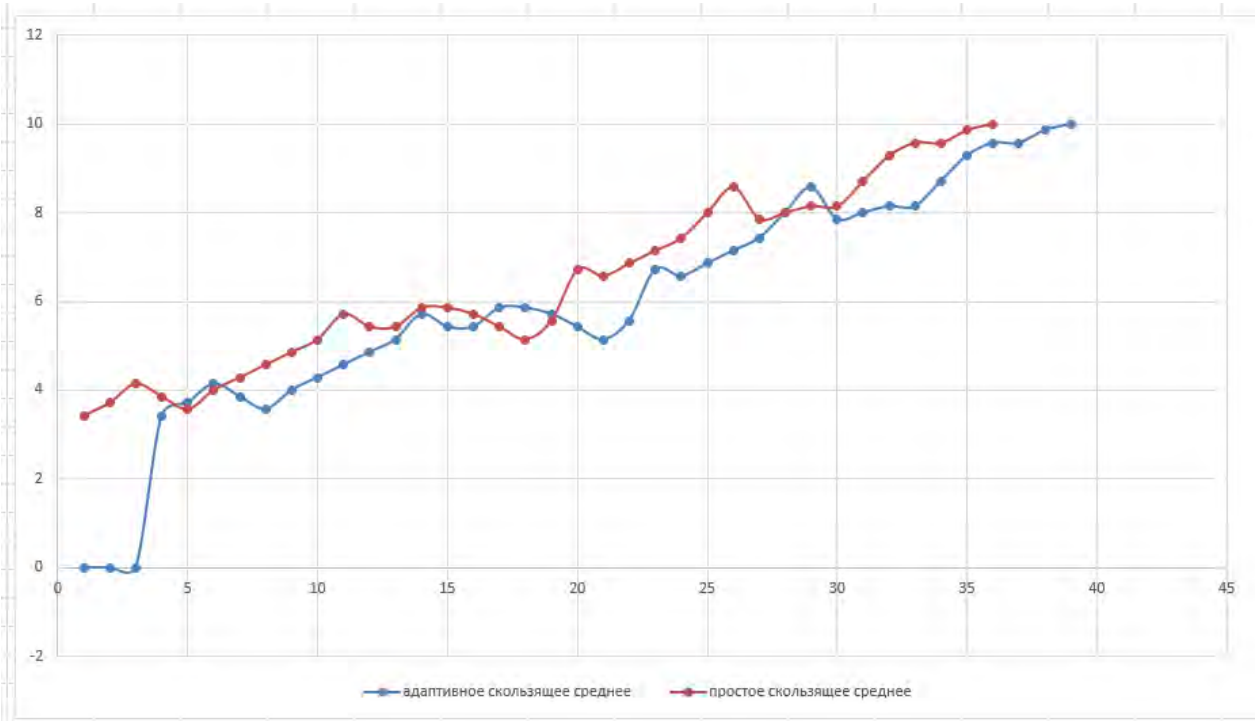


Рис. 9.2 – Сравнение простого $\bar{y}_i$ и адаптивного $\tilde{y}_i$ скользящих средних

На рис. 9.3 в столбце B приведены результаты сглаживания временного ряда радиопередач с помощью процедуры *Экспоненциальное сглажи*вание (параметр сглаживания $\alpha = 0{,}1$, фактор затухания $\beta = 0{,}9$).

	A	B	C
1	1	#Н/Д	1
2	2	1	1,10
3	8	1,10	1,79
4	10	1,79	2,61
5	2	2,61	2,55
6	1	2,55	2,39
7	0	2,39	2,16
8	3	2,16	2,24
9	5	2,24	2,52
10	6	2,52	2,86
11	8	2,86	3,38
12	5	3,38	3,54
13	3	3,54	3,49
14	2	3,49	3,34
15	5	3,34	3,50
16	7	3,50	3,85
17	10	3,85	4,47
18	6	4,47	4,62
19	5	4,62	4,66
20	6	4,66	4,79
21	2	4,79	4,51
22	4	4,51	4,46
23	5	4,46	4,52
24	8	4,52	4,86
25	9	4,86	5,28

Рис. 9.3 – Решение примера 9.1 (экспоненциальное сглаживание)

На этом же рисунке в столбце C приведены результаты «канонического» экспоненциального сглаживания (вычисления выполнены по формуле $= 0{,}9 * C1 + 0{,}1 * A2$, введенной в ячейку C2 и скопированной затем в ячейки C3:C42). На рис. 9.4 для сравнения приведены график экспоненциального сглаживания, сформированный процедурой, и график, построенный по дан-

ным, хранящимся в диапазоне C2:C42 (этот график имеет пометку «Каноническая формула»).

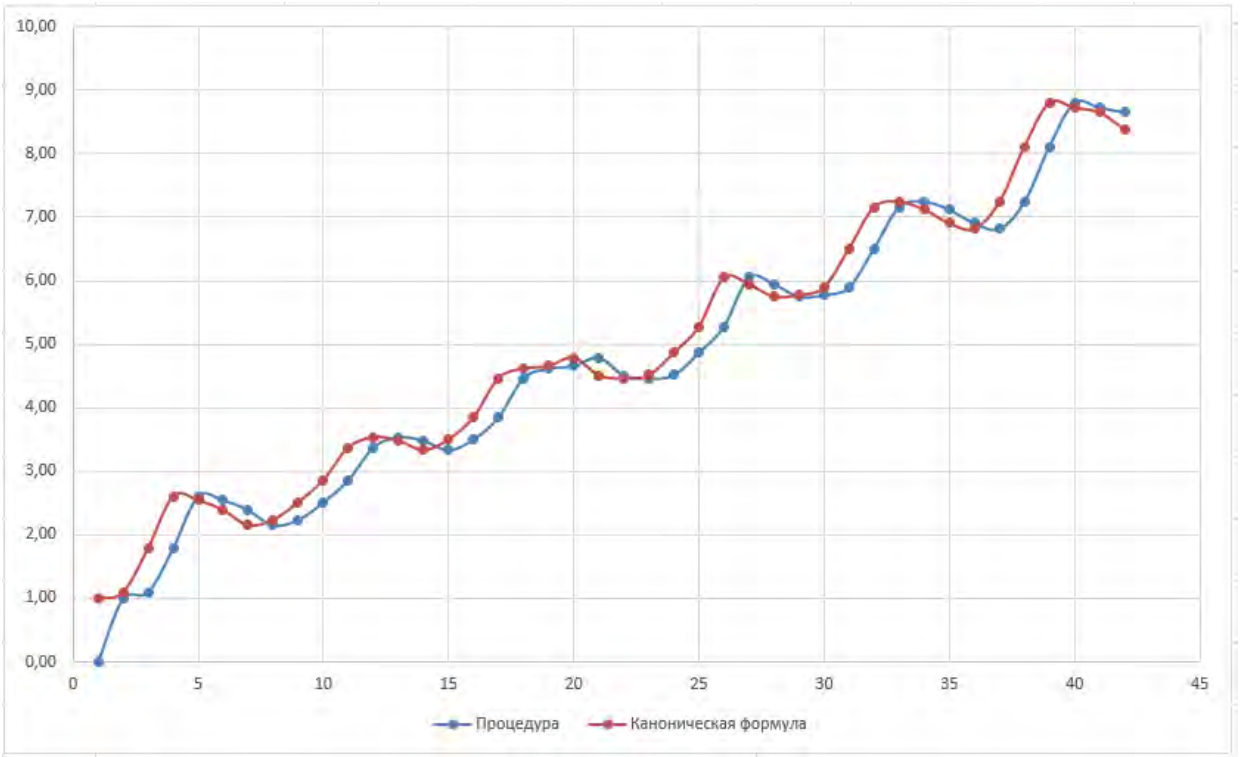


Рис. 9.4 – Сравнение «канонического» экспоненциального сглаживания и сглаживания, выполненного с помощью процедуры *Экспоненциальное сглаживание*

Контрольный пример 9.2. Имеются данные об объемах продаж некоторой фирмы (табл. 9.2). С помощью графика подобрать линию тренда, которая лучше всего описывает фактические данные и на ее основе сделать прогноз на три недели вперед.

Таблица 9.2

Неделя	1	2	3	4	5	6	7	8	9	10	11
Количество продаж	17	22	26	27	35	40	41	45	50	63	78

Решение. В ячейки A1 и B1 введем заголовки исходных данных, в ячейки A2:A12 – номера недель, а в ячейки B2:B12 – соответствующее количество продаж (фактические данные). По этим данным построим диаграмму фактических значений показателя (рис. 9.5).

	A	B
1	Неделя	Количество продаж
2	1	17
3	2	22
4	3	26
5	4	27
6	5	35
7	6	40
8	7	41
9	8	45
10	9	50
11	10	63
12	11	78

Рис. 9.5 – Исходные данные и график фактических значений показателя.

Выделим рад данных щелчком по любой точке ряда, вызовем контекстное меню (правой кнопкой мыши) и выберем в нем команду *Добавить линию тренда*. На экране появляется окно *Формат линии тренда*.

Выбираем параметры линии тренда – *Линейная* и устанавливаем флажки:

- показывать уравнение на диаграмме;
- поместить на диаграмму величину достоверности аппроксимации (R^2).

После нажатия кнопки *Закрыть* на графике наряду с фактическими значениями количества продаж будет показана линейная функция тренда и ее уравнение (рис. 9.6). Уравнение и коэффициент детерминации можно выделить щелчком левой кнопки мыши и перетащить ее на то место графика, где их лучше видно.

Рис. 9.6 – Линейная кривая роста и ее уравнение

Аналогично следует попробовать другие типы линии тренда. При добавлении каждой новой линии на график нужно сравнить ее коэффициент детерминации с аналогичным показателем предыдущей модели.

MS Excel дает возможность добавлять на график полиномы до 6-й степени включительно. Но чем больше степень полинома (т.е. больше параметров), тем больше должно быть исходных данных. Поскольку мы рассматриваем не так много уровней временного ряда, ограничимся полиномом второй степени.

В таблицу 9.3 занесем линию тренда и соответствующее значение R^2.

Таблица 9.3

№	Линия тренда	R^2
1	$y = 5{,}3x + 8{,}5636$	$R^2 = 0{,}925$
2	$y = 21{,}273 \ln x + 6{,}5161$	$R^2 = 0{,}7521$
3	$y = 26{,}257e^{0{,}1361x}$	$R^2 = 0{,}976$
4	$y = 14{,}483x^{0{,}5858}$	$R^2 = 0{,}9129$
5	$y = 0{,}399x^2 + 0{,}5028x + 18{,}958$	$R^2 = 0{,}966$

Сравнивая величину коэффициента детерминации R^2, в качестве «наилучшего» приближения выбираем экспоненциальную модель, поскольку для нее коэффициент детерминации наибольший (рис. 9.7).

Рис. 9.7 – Экспоненциальная линия тренда,
наиболее точно описывающая исходные данные задачи

Так как нужно выполнить прогноз на 3 недели вперед, допишем номера этих недель (12, 13 и 14) в столбец А. Столбец С озаглавим «Теоретические значения» и занесем в него формулы расчета по выбранной функции тренда. В ячейку С2 запишем формулу = 16,257*EXP(0,1361*A2). Эта формула копируется методом автозаполнения в ячейки С3:С15, т.е. теоретические значения рассчитываются для всех моментов времени в прошлом и прогнозируемом будущем (рис. 9.8).

В результате получим в ячейках С13:С15 следующие точечные прогнозы:

- на 12-ю неделю – 83 продажи;
- на 13-ю неделю – 95 продаж;
- на 14-ю неделю – 109 продаж.

	A	B	C	D
1	**Неделя**	**Количество продаж**	**Теоретические значения**	
2	**1**	**17**	**19**	
3	**2**	**22**	**21**	
4	**3**	**26**	**24**	
5	**4**	**27**	**28**	
6	**5**	**35**	**32**	
7	**6**	**40**	**37**	
8	**7**	**41**	**42**	
9	**8**	**45**	**48**	
10	**9**	**50**	**55**	
11	**10**	**63**	**63**	
12	**11**	**78**	**73**	
13	**12**		**83**	
14	**13**		**95**	
15	**14**		**109**	

Рис. 9.8 – Прогнозирование продаж на три недели вперед

Контрольный пример 9.3. Построить регрессионную модель для предсказания изменений уровня заболеваемости органов дыхания (Y) в зависимости от содержания в воздухе двуокиси углерода (X_1) и степени запыленности (X_2). В таблице 9.4 приведены данные наблюдений в течение 28 месяцев.

Предсказать уровень заболеваемости при содержании двуокиси углерода, равной 0,7, и запыленности 1,5.

Задание выполнить в пакетах *Excel* и *Statistica*.

Таблица 9.4

№	1	2	3	4	5	6	7	8	9	10
X_1	1	1	1.1	1.1	1.1	1.1	1	1	1.2	1.2
X_2	1.3	1.3	1.4	1.4	1.5	1.5	1.4	1.5	1.6	1.7
Y	1160	1155	1158	1157	1160	1161	1157	1159	1256	1260
№	11	12	13	14	15	16	17	18	19	20
X_1	0.6	0.6	0.7	0.7	0.75	0.7	0.7	0.7	0.8	0.8
X_2	1	1	1.1	1.15	1.2	1.2	1.3	1.3	1.4	1.4
Y	1040	1039	1039	1040	1040	1039	1040	1039	1140	1138
№	21	22	23	24	25	26	27	28		
X_1	0.78	0.8	0.78	0.78	0.8	0.8	0.75	0.78		
X_2	1.5	1.5	1.5	1.6	1.7	1.8	1.8	1.9		
Y	1240	1239	1241	1240	1239	1239	1240	1238		

Решение. Введем исходные данные, расположив каждую случайную величину в отдельном столбце (на рис. 9.9 показаны первые 14 строк исходных данных).

	A	B	C	D	E	F
1	Содержание CO2 (X1)	Запыленность (X2)	Уровень заболеваемости (Y)		R(X1Y)	0,475038
2	1	1,3	1160		R(X2Y)	0,87521
3	1	1,3	1155			
4	1,1	1,4	1158			
5	1,1	1,4	1157			
6	1,1	1,5	1160			
7	1,1	1,5	1161			
8	1	1,4	1157			
9	1	1,5	1159			
10	1,2	1,6	1256			
11	1,2	1,7	1260			
12	0,6	1	1040			
13	0,6	1	1039			
14	0,7	1,1	1039			
15	0,7	1,15	1040			

Рис. 9.9 – Исходные данные для регрессионной модели

Оценим наличие и силу линейных связей между зависимой величиной Y и каждым из факторов X_1 и X_2. Для этого рассчитаем коэффициенты линейной корреляции, введя в ячейки F1 и F2 формулы = КОРРЕЛ(A2:A29; C2:C29) и = КОРРЕЛ(B2:B29; C2:C29). Получим $r_{X1Y} = 0{,}475$ и $r_{X2Y} = 0{,}875$. Таким образом, связь между заболеваемостью и содержанием CO_2 является умеренной, а между заболеваемостью и запыленностью высокой.

Построим двухфакторную регрессионную модель вида $\tilde{y} = a_0 + a_1 X_1 + a_2 X_2$. В меню *Данные* выберем *Анализ данных – Регрессия*. Заполним диалоговое окно, как показано ни рис. 9.10.

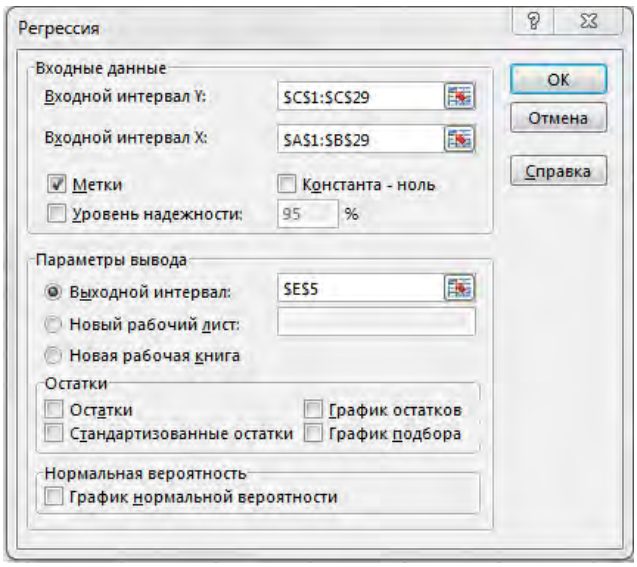


Рис. 9.10 – Пример заполнения диалогового окна *Регрессия*

Результаты работы процедуры Регрессия из Пакета Анализа показаны на рис. 9.11.

ВЫВОД ИТОГОВ

Регрессионная статистика	
Множественный R	0,8897
R-квадрат	0,7916
Нормированный R-квадрат	0,7750
Стандартная ошибка	39,3332
Наблюдения	28

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	2	146954,6596	73477,32982	47,49365234	3,05534E-09
Остаток	25	38677,44751	1547,0979		
Итого	27	185632,1071			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
Y-пересечение	672,4988584	51,26494841	13,11810271	1,0396E-12	566,9167208	778,0809961
Содержание CO2 (X1)	79,69364308	45,42450365	1,754419678	0,091610092	-13,85987344	173,2471596
Запыленность (X2)	288,8816344	35,05502985	8,240804118	1,36259E-08	216,6844489	361,0788198

Рис. 9.11 Результаты регрессионного анализа

Значения коэффициентов регрессии находятся в столбце *Коэффициенты* и соответствуют:

- Y-пересечение – a_0;
- Содержание CO_2 – a_1;
- Запыленность – a_2.

Таким образом, получаем следующее уравнение регрессии:

$$\tilde{y} = 672{,}5 + 79{,}7X_1 + 288{,}9X_2.$$

Для каждого коэффициента рассчитана также стандартная ошибка и выборочное значение *t*-статистики (отношение оценки параметра к ее стандартной ошибке). Для оценки достоверности отличия каждого параметра от нуля найдем критическое значение критерия Стьюдента для уровня значимости $\alpha = 0{,}05$. Введем в произвольную ячейку формулу =СТЬЮДЕНТ.ОБР.2Х(0,05;26) – здесь $26 = n - 2$:

H	I	J
	2,056	

Сравнивая это значение с *t*-статистикой для каждого параметра, убеждаемся, что для a_0 и a_2 выполняется условие $t > t_{кр}$ (13,118 > 2,056 и 8,251 > 2,056), а для a_1 – нет (1,754 < 2,056). Поэтому параметры a_0 и a_2 можно считать достоверно отличны от нуля, а параметр a_1 – нельзя.

Аналогично результаты дает столбец *P – значение* для гипотезы о равенстве параметра нулю. Так как для a_0 и a_2 эта вероятность значительно меньше уровня значимости $\alpha = 0{,}05$ ($1{,}039 \cdot 10^{-12} < 0{,}05$ и $1{,}362 \cdot 10^{-8} < 0{,}05$), нулевая гипотеза отклоняется. Для a_1 вероятность принятия нулевой гипотезы получилась немного больше, чем уровень значимости $(0{,}09 > 0{,}05)$, что не дает права отвергнуть ее. Таким образом, фактор содержания CO_2 нуждается в дополнительном исследовании. Возможно, его влияние на заболеваемость носит нелинейный характер. Возможно также, что фактических данных недостаточно для доказательства его влияния.

Точность регрессионной модели оценивается на основании коэффициента детерминации $R^2 = 0{,}7916$ (соответствующая строка в таблице *Регрессионная статистика*). Поскольку это значение близко к 0,8, можно говорить о том, что точность модели удовлетворительная.

Рассчитаем теперь прогноз заболеваемости, подставив в уравнение регрессии заданные в условии значения X_1 и X_2. Для этого занесем эти значения в ячейки Excel, как показано на рис. 9.12. В ячейку J4 запишем формулу уравнения регрессии, причем в качестве параметров можно использовать ссылки на соответствующие ячейки выходного диапазона. Результат расчета по этой формуле (прогнозные значения заболеваемости) составляет приблизительно 1162.

H	I	J
	=СТЬЮДЕНТ.ОБР.2Х(0,05;26)	
	Прогноз	
X1	X2	Y
0,7	1,5	=$F$21+$F$22*H4+$F$23*I4

Рис. 9.12 – Расчет прогноза заболеваемости

Решим данную задачу с применением пакета *Statistica*. Образуем таблицу с 3 столбцами и 28 строками. Вводим в таблицу исходные данные (рис.9.13):

	1 X1	2 X2	3 Y
1	1	1,3	1160
2	1	1,3	1155
3	1,1	1,4	1158
4	1,1	1,4	1157
5	1,1	1,5	1160
6	1,1	1,5	1161
7	1	1,4	1157
8	1	1,5	1159
9	1,2	1,6	1256
10	1,2	1,7	1260
11	0,6	1	1040
12	0,6	1	1039
13	0,7	1,1	1039
14	0,7	1,15	1040
15	0,75	1,2	1040
16	0,7	1,2	1039
17	0,7	1,3	1040
18	0,7	1,3	1039
19	0,8	1,4	1140
20	0,8	1,4	1138
21	0,78	1,5	1240
22	0,8	1,5	1239
23	0,78	1,5	1241
24	0,78	1,6	1240
25	0,8	1,7	1239
26	0,8	1,8	1239
27	0,75	1,8	1240
28	0,78	1,9	1238

Рис. 9.13 – Исходные данные задачи

Вызовем модуль *Множественная регрессия (Multiple Regression)* и укажем имена рядов с помощью кнопки *Переменные (Variables)*: одного зависимого (*dependent*) и нескольких независимых *(independent*) (рис. 9.14).

Далее надо дать модулю указание в поле *Файл данных* (*Input file*) относительно вида исходных данных. Модуль одинаково успешно работает с данными, заданными как в виде таблицы наблюдений *Исходные данные* (*Raw Data*), так и в виде корреляционной матрицы (*Correlation Matrix*). Выберем *Raw Data* (рис. 9.14).

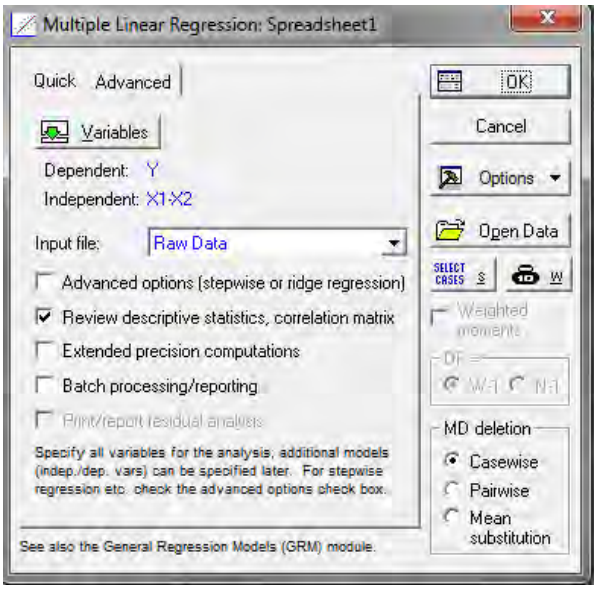


Рис. 9.14 – Стартовая панель модуля множественной регрессии

Если есть необходимость посмотреть матрицу коэффициентов корреляции и описательные статистики рядов, то надо поставить галочку у *Описательной статистики. (Review descriptive statistics*) в стартовом окне модуля.

Появится промежуточное окно с названием *Просмотр описательных статистик* (*Review Descriptive Statistics*), где можно кнопкой *Среднее и стандартное отклонение* (*Mean & Standard deviation*) получить значения основных числовых характеристик факторов (рис. 9.15 – 1), а кнопкой *Корреляция* (*Correlation*) получить матрицу коэффициентов корреляции (рис. 9.15 – 2).

Variable	Means and Standard Dev		
	Means	Std.Dev.	N
X1	0,861	0,17960	28
X2	1,427	0,23273	28
Y	1153,321	82,91721	28

1

Variable	Correlations (Spreadsheet1)		
	X1	X2	Y
X1	1,000000	0,372974	0,475038
X2	0,372974	1,000000	0,875210
Y	0,475038	0,875210	1,000000

2

Рис. 9.15 – Просмотр описательных статистик

В *модуле* имеются на выбор три модели множественной регрессии: 1) стандартная; 2) с автоматическим включением новых предикторов; 3) с автоматическим исключением предикторов из заданного набора.

Стандартная модель работает «по умолчанию» и сразу выводит окно с результатами подбора уравнения регрессии. Но если нужно использовать модели с автоматическим включением или исключением предикторов, то необходимо поставить галочку у *Пошаговая и гребневая регрессия* (*Advanced Options*) (рис. 9.14).

В окне *Review Descriptive Statistics* нажмем кнопку ОК.

Диалоговое окно результатов (*Multiple Regression Results)* состоит из двух частей – информационной и функциональной (рис. 9.16).

В информационной (верхней) части окна результатов важными являются значение множественной корреляции R и коэффициента детерминации R^2. Анализировать надо скорректированное значение R^{*2} (*adjusted R2*), представляющее собой его несмещённую оценку. Если R^{*2} мало, значит, не учтено влияние каких-то факторов, существенных для процесса формирования Y.

В нашем случае коэффициент детерминации $R^{*2} = 0{,}775$.

Считается, что набор предикторов достаточно хорошо отражает условия формирования переменной Y, если $R^{*2} \geq 0{,}5$.

Здесь же приводится F-статистика Фишера проверки адекватности модели данным. Проверяется нулевая гипотеза о равенстве нулю коэффициентов уравнения регрессии. По приведённому в окне уровню значимости *p*, со-

ответствующему F-статистике, можно сделать предварительное заключение о пригодности уравнения. Если уровень значимости $p \le p$ крит (по умолчанию p крит $= 0{,}05$), то уравнение регрессии с данным набором предикторов пользоваться не имеет смысла.

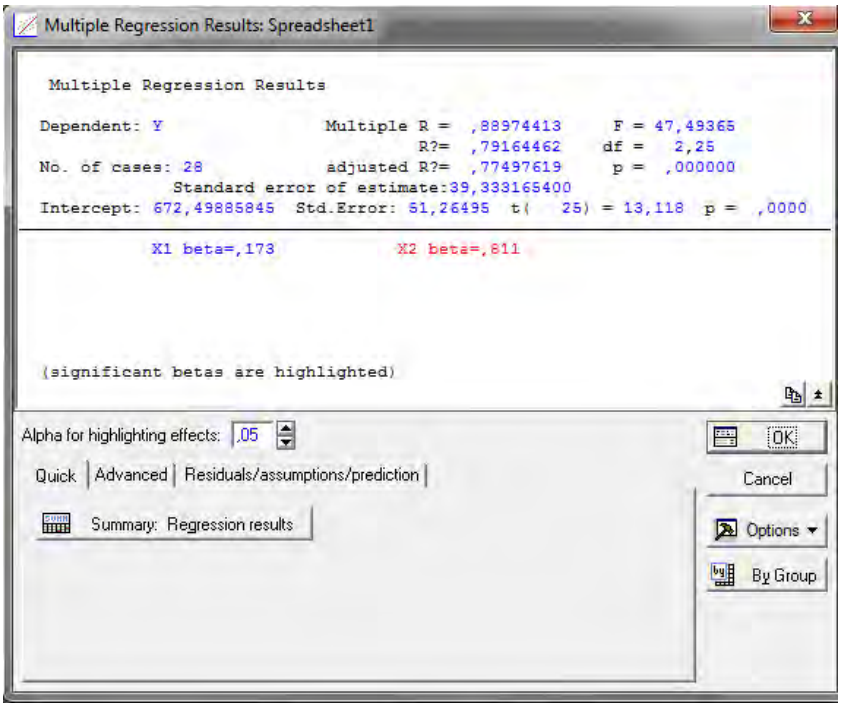


Рис. 9.16 – Окно результатов

Важную информацию несут *beta*-коэффициенты. Это стандартизованные значения коэффициентов уравнения регрессии. Преимущество *beta* коэффициентов в том, что они позволяют сравнивать относительный вклад каждой независимой переменной в прогнозе. Их интерпретация подобна анализу частных коэффициентов корреляции. Например, согласно информации, в окне результатов рисунка 9.16, фактор X2 вносит в прогноз величины *Y* значительно больший вклад, чем X1. Значимые *beta* -коэффициенты выделяются красным цветом. Факторы, имеющие незначимые *beta*-коэффициенты, из уравнения регрессии удаляются как неинформативные.

Нажав на кнопку *Summary: Regression results*, получим таблицу результатов (рис.9.17):

Regression Summary for Dependent Variable: Y (Spreadsheet1)
R= ,88974413 R?= ,79164462 Adjusted R?= ,77497619
F(2,25)=47,494 p<,00000 Std.Error of estimate: 39,333

N=28	Beta	Std.Err. of Beta	B	Std.Err. of B	t(25)	p-level
Intercept			672,4989	51,26495	13,11810	0,000000
X1	0,172620	0,098392	79,6936	45,42450	1,75442	0,091610
X2	0,810827	0,098392	288,8816	35,05503	8,24080	0,000000

Рис. 9.17 – Результаты регрессионного анализа

Обобщённый коэффициент корреляции равен: $|R| = 0{,}8897$. Остаточная дисперсия – $D_{\text{ост}} = 39{,}33^2 = 1548{,}8$. В столбце В указаны оценки неизвестных коэффициентов. Таким образом, оценка неизвестной функции регрессии имеет вид $\widetilde{y} = 672{,}5 + 79{,}7X_1 + 288{,}9X_2$.

Адекватность модели данным доказывается анализом остатков.

Чтобы уравнением регрессии можно было пользоваться на практике, нужно показать, что остатки независимы и распределены по нормальному закону.

В модуле для проверки независимости остатков используется статистика Дарбина – Уотсона, являющаяся стандартным методом обнаружения их автокоррелированности.

Статистика *d* Дарбина – Уотсона используется для проверки гипотезы о том, что остатки построенной регрессионной модели некоррелированы (корреляции равны нулю), против альтернативы: остатки связаны авторегрессионной зависимостью. Вычисленное значение статистики *d* надо сравнить с двумя критическими: нижним DW_1 и верхним DW_2.

- Если $d < DW_1$ (или $4 - d < DW_1$), то в остатках имеется автокорреляция на заданном уровне значимости.
- Если $d > DW_2$ (или $4 - d > DW_2$), то автокорреляция отсутствует.
- Если $DW_1 < d < DW_2$, то случай сомнительный, нужны дополнительные исследования.

Когда расчётное значение статистики *d* превышает 2, то с DW_1 и DW_2 сравнивается не сам коэффициент d, а выражение $4 - d$.

Критические точки для данного числа наблюдений и числа факторов находят в таблице, составленной для определенного уровня значимости. В таблице 9.5 приводятся значения критических точек статистики Дарбина-Уотсона для уровня значимости $\alpha = 0{,}05$.

Таблица 9.5

Таблица критических точек статистики Дарбина – Уотсона

(n – число наблюдений, m – число факторов, α = 0,05)

n	$m = 1$		$m = 2$		$m = 3$		$m = 4$		$m = 5$	
	DW_1	DW_2	DW_1	DW_2	DW_1	DW_2	DW_1	DW_2	DW_1	DW_2
15	1,08	1,36	0,95	1,54	0,82	1,75	0,69	1,97	0,56	2,21
16	1,1	1,37	0,96	1,54	0,86	1,73	0,74	1,93	0,62	2,15
17	1,03	1,38	1,02	1,54	1,9	1,71	0,78	1,9	0,67	2,1
18	1,16	1,39	1,05	1,53	0,93	1,69	0,82	1,87	0,71	2,06
19	1,18	1,4	1,08	1,53	0,97	1,68	0,86	1,85	0,75	2,02
20	1,2	1,41	1,1	1,54	1	1,68	0,9	1,83	0,79	1,99
21	1,22	1,42	1,13	1,54	1,03	1,67	0,93	1,81	0,83	1,96
22	1,24	1,43	1,15	1,54	1,05	1,66	0,98	1,8	0,86	1,94
23	1,26	1,44	1,17	1,54	1,08	1,66	0,99	1,79	0,9	1,92
24	1,27	1,45	1,19	1,55	1,1	1,66	1,01	1,78	0,93	1,9
25	1,29	1,45	1,21	1,55	1,12	1,66	1,04	1,77	0,95	1,89
26	1,3	1,46	1,22	1,56	1,14	1,65	1,06	1,76	0,98	1,88
27	1,32	1,47	1,24	1,56	1,16	1,65	1,08	1,76	1,01	1,86
28	1,33	1,48	1,26	1,56	1,18	1,65	1,1	1,76	1,03	1,85
29	1,34	1,48	1,27	1,56	1,2	1,65	1,12	1,74	1,05	1,84
30	1,35	1,49	1,28	1,57	1,21	1,65	1,14	1,74	1,07	1,83

Вычислим статистику Дарбина-Уотсона для остатков в диалоговом окне *Анализ остатков* (*Residual Analysis*) на вкладке *Дополнительно* (*Advanced*) (рис. 9.18, 9.19).

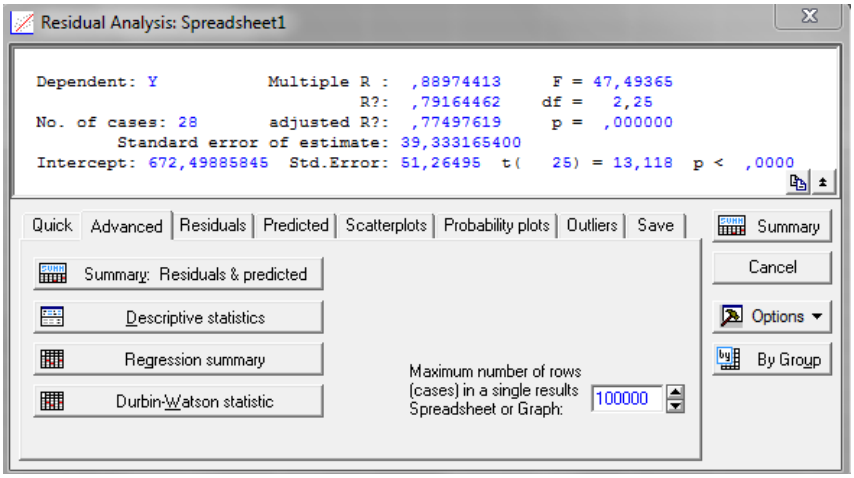


Рис. 9.18 – Окно анализа остатков

С помощью таблицы критических точек статистики Дарбина-Уотсона выясним, являются ли остатки независимыми.

	Durbin-Watson d (Spreadsheet1) and serial correlation of residuals	
	Durbin-Watson d	Serial Corr.
Estimate	0,623805	0,684535

Рис. 9.19 – Окно с вычисленной статистикой Дарбина -Уотсона

Так как $DW_1 = 1{,}26$; $DW_2 = 1{,}56$ и $d < DW_1$, то гипотеза о независимости случайных отклонений отвергается – присутствует положительная автокорреляция.

Для прогноза в окне результатов на вкладке *Остатки / предсказанные / наблюденные значения* (*Residials/ assumptions/ predictions*) имеется кнопка с зеленой стрелкой *Предсказать зависимую переменную* (

Она вызывает специальное окно для прогноза (рис. 9.20), в котором надо задать числовые значения факторов X_i.

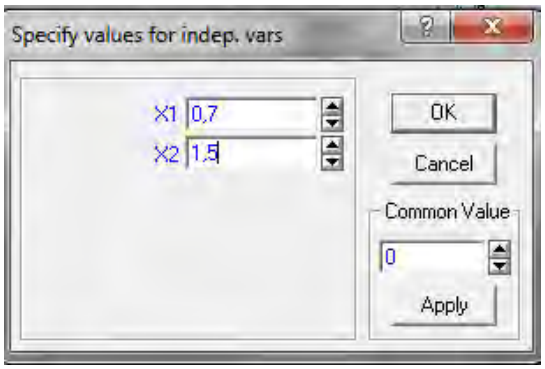


Рис. 9.20 – Окно задания прогноза

Результаты работы появятся в виде маленькой таблицы, в нижнем правом углу которой будет напечатано спрогнозированная по уравнению регрессии величина (1161.607) и её доверительные 95% интервалы (рис. 9.21).

Variable	Predicting Values for (Spreadsheet1) variable: Y		
	B-Weight	Value	B-Weight * Value
X1	79,6936	0,700000	55,786
X2	288,8816	1,500000	433,322
Intercept			672,499
Predicted			1161,607
-95,0%CL			1138,156
+95,0%CL			1185,058

Рис. 9.21 Окно с результатами прогноза по уравнению регрессии

Задания для самостоятельной работы

Задача 1. Для данного временного ряда выполнить выравнивание (в пакете *Excel*):

а) методом скользящих средних (интервал m); б) методом экспоненциального сглаживания ($\beta = 1 - \alpha$).

Построить графики фактических и прогнозных значений. Визуально определить, какое из значений α наиболее соответствует процессу, заданному временным рядом.

Вариант 1 $m = 3;\ \alpha = 0{,}2$										
t	1	2	3	4	5	6	7	8	9	
x	26,75	20,75	20,9	20,55	19,33	26,27	23,86	22,69	21,85	
t	10	11	12							
x	17,95	32,56	43,91							
Вариант 2 $m = 5;\ \alpha = 0{,}3$										
t	1	2	3	4	5	6	7	8	9	
x	3000	3017	3018	3005	2990	2975	2973	2996	3010	
Вариант 3 $m = 4;\ \alpha = 0{,}4$										
t	1	2	3	4	5	6	7	8	9	10
x	67	59	47	62	64	46	51	37	63	56
Вариант 4 $m = 3;\ \alpha = 0{,}3$										
t	1	2	3	4	5	6	7	8	9	10
x	212,1	215	216,6	219,2	221,4	220,7	220,5	222,6	224,8	228,2
Вариант 5 $m = 3;\ \alpha = 0{,}3$										
t	1	2	3	4	5	6	7	8	9	10
x	4633	4742	4980	5187	5107	5358	5299	5228	5345	5244
Вариант 6 $m = 3;\ \alpha = 0{,}3$										
t	1	2	3	4	5	6	7	8	9	10
x	7,9	8,3	7,5	6,9	7,2	5,6	5,8	4,9	5,1	4,4
Вариант 7 $m = 5;\ \alpha = 0{,}2$										
t	1	2	3	4	5	6	7	8		
x	3946,4	4058,8	4121,2	4102,6	4102,1	4159,2	4185,8	4250,1		
Вариант 8 $m = 5;\ \alpha = 0{,}4$										
t	1	2	3	4	5	6	7	8		
x	7,4	6,8	6,1	5,6	5,4	4,9	4,5	4,2		
Вариант 9 $m = 3;\ \alpha = 0{,}2$										
t	1	2	3	4	5	6	7	8	9	
x	171	147	169	162	186	181	168	222	195	

Вариант 10 $m=3;\ \alpha=0{,}3$									
t	1	2	3	4	5	6	7	8	9
x	111.8	128.2	111.5	110.3	111.5	120	117.4	114.8	119.2
t	10	11							
x	103.2	117.2							
Вариант 11 $m=3;\ \alpha=0{,}2$									
t	1	2	3	4	5	6	7	8	9
x	33	40	28	37	30	44	36	32	40
t	10	11	12						
x	37	49	36						
Вариант 12 $m=3;\ \alpha=0{,}2$									
t	1	2	3	4	5	6	7	8	9
x	6	4.4	5	9	7.2	4.8	6	10	8
t	10	11	12	13	14	15	16		
x	5.6	6.4	11	9	6.6	7	10.8		

Задача 2. Имеются данные об объемах продаж некоторой фирмы. С помощью графика подобрать линию тренда, которая лучше всего описывает фактические данные и на ее основе сделать прогноз на 3 недели вперед. Задание выполнить в пакете *Excel*.

Вариант 1												
Неделя	1	2	3	4	5	6	7	8	9			
Количество продаж	3	10	11	18	21	30	32	37	42			
Вариант 2												
Неделя	1	2	3	4	5	6	7	8	9	10	11	12
Количество продаж	8	8	9	12	17	16	26	25	35	33	39	52
Вариант 3												
Неделя	1	2	3	4	5	6	7	8	9	10	11	
Количество продаж	40	42	43	46	49	50	56	62	65	63	70	
Вариант 4												
Неделя	1	2	3	4	5	6	7	8	9	10	11	
Количество продаж	18	25	27	26	30	37	35	40	48	53	57	
Вариант 5												
Неделя	1	2	3	4	5	6	7	8	9	10	11	12
Количество продаж	6.75	6.24	5.98	5.76	4.9	4.57	4.32	4.02	3.78	3.54	3.36	2.28
Вариант 6												
Неделя	1	2	3	4	5	6	7	8	9			
Количество продаж	28.3	24.4	25	28.9	38.3	54.4	64.4	72.7	97.7			

Вариант 7												
Неделя	1	2	3	4	5	6	7	8	9	10	11	
Количество продаж	17	22	26	27	35	40	41	45	50	63	78	
Вариант 8												
Неделя	1	2	3	4	5	6	7	8	9	10	11	12
Количество продаж	45	65	85	99	102	101	120	107	120	115	118	121
Вариант 9												
Неделя	1	2	3	4	5	6	7	8	9	10	11	12
Количество продаж	239	201	182	297	324	278	257	384	401	360	335	462
Вариант 10												
Неделя	1	2	3	4	5	6	7	8	9	10		
Количество продаж	15	15	16	22	32	39	49	58	65	67		
Вариант 11												
Неделя	1	2	3	4	5	6	7					
Количество продаж	659	656.7	645.1	633.6	621.7	605.6	585.03					
Вариант 12												
Неделя	1	2	3	4	5	6	7	8	9	10	11	12
Количество продаж	70	66	65	71	79	66	67	82	84	69	72	87

Задание 3. По представленным данным аппроксимировать статистическую зависимость величины Y от X_1 и X_2 функцией $\widetilde{y} = a_0 + a_1 X_1 + a_2 X_2$. Вычислить остаточную дисперсию, найти оценку обобщённого коэффициента корреляции. В пакете *Statistica* провести анализ остатков (проверить адекватность модели данным).

Значения аргументов X_1 и X_2 представлены ниже

X_1	1,1	1,2	1,3	1,4	1,5	1,6	1,7	1,8	2,1	2,2	2,3	2,4	2,5	2,6	2,7
X_2	2,1	3,5	2,6	1,3	2,2	1,8	1,5	1,9	1,4	2,5	1,7	2,1	1,6	1,5	2,3

Соответствующие значения функции Y выбрать по номеру варианта.

1	1,1	5,9	3,4	–1,1	2,2	1,2	0,4	1,8	0,5	4,4	2,1	3,6	3	1,9	4,7
2	1,4	4,9	2,8	–0,8	2,3	0,8	0,2	1,8	0,7	4,2	2,2	3,4	2,4	2,4	5,2
3	0,9	5,6	2,6	–0,5	2,2	1,1	0,5	2	0,4	4,8	2,1	3,9	2,3	2,3	4,4
4	1	5,4	3,4	–0,9	2,4	1,2	0,5	1,8	0,6	3,9	2,1	3,3	2	2,4	5,3
5	0,9	5,8	3,2	–0,5	1,8	0,7	0,1	1,6	1,1	4,9	2,2	3,9	2,5	1,9	5,3
6	5,7	7,5	6,3	5,8	6,6	6,7	6,7	6,9	7,2	8	7,4	8,2	7,9	8,3	9,2
7	5,9	7,1	6,8	5,5	7,1	6	6,6	7,5	7	8,5	8,1	8,1	8	7,8	8,8

8	6,3	7,7	6,6	5,9	6,4	6,5	6,4	7	6,8	8,7	7,5	8	8,3	8,1	9,3
9	6,2	7,2	6,8	6	6,3	6,9	6,7	7,3	7,4	8,4	7,4	8,1	8,7	7,8	9,7
10	5,7	7,2	7	5,9	6,2	6,8	6,8	7	7,3	8,9	8,2	8,4	7,9	8,1	9,2
11	4,9	7,6	5,9	3,5	5,3	4,9	4,6	4,6	4,8	6,9	5,5	5,7	5,8	5	7,1
12	5,1	7,5	6,2	3,6	5,8	5	4,7	5,4	4	6,2	5	6,5	5	4,6	7,1

2.2. БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Лисьев, В П. Теория вероятностей и математическая статистика: учебное пособие/ Московский государственный университет экономики, статистики и информатики. – М., 2006. – 199 с.
2. Маталыцкий, М.А. Теория вероятностей и математическая статистика: пособие / М.А. Маталыцкий, Т.В. Русилко. – 2-е изд., перераб. и доп.// Гродно: ГрГУ, 2009. – 219 с.
3. Математическая статистика в примерах и задачах: учебное пособие / Г.С. Евдокимова, Смол. гос. ун-т. – Смоленск: Изд-во СмолГУ, 2014. 98 с. – Режим доступа https://studfile.net/preview/3544158/.
4. Гмурман, В.Е.. Теория вероятностей и математическая статистика/ В.Е. Гмурман. – М., Высшая школа, 2003. – 479 с.
5. Вадзинский, Р. Статистические вычисления в среде Excel. Библиотека пользователя. – СПб, Питер, 2008. – 608 с., ил.
6. Аверьянова, С. Ю. Лабораторный практикум по математической статистике в среде ЭТ MS Excel: учебное пособие/ С.Ю. Аверьянова, Н.В. Растеряев// Южный федеральный университет. – Ростов-на-дону: Издательство Южного федерального университета, 2014. – 64 с.
7. Еськова, О.И. Основы статистической обработки информации: пособие/ О.И. Еськова, Л.П. Авдашкова, М.А. Грибовская. – Минск: Беларусь, 2011. – 175 с.: ил.
8. Математическая статистика: учеб-метод. пособие / авт-сост.: С.Е. Демин, Е.Л. Демина; М-во образования и науки РФ; ФГАОУ ВО «УрФУ им. Первого президента России Б.Н. Ельцина», нижнетагил. технол. ин-т (фил.). – Нижний Тагил: НТИ (филиал) УрФУ, 2016. – 284 с.
9. Решение задач теории вероятностей, математической статистики и математического программирования средствами EXCEL: метод. указания к выполнению самост. работы для бакалавров техн. и экон. направлений подготовки очной формы обучения / сост: Ю.Г. Кошкин, Е.П. Погодина, Н.Г. Тетерина; Сиб. гос. аэрокосмич. ун-т. – Красноярск, 2014 – 88 с.
10. Лунгу, К.Н. Высшая математика. Руководство к решению задач. Ч.2 / К.Н. Лунгу, Е.В. Макаров – М.: ФИЗМАТЛИТ, 2007. – 384 с.
11. Стукач, О.В. Программный комплекс Statistica в решении задач управления качеством: учебное пособие / О.В. Стукач; Томский политехни-

ческий университет. – Томск: изд-во Томского политехнического университета, 2011. – 163 с.
12. Дубровина, О.В. Прикладная математика: метод. пособие по выполнению практических и лабораторных работ для студентов заочного отделения специальности 1-54 01 01 «Метрология, стандартизация и сертификация» / О.В. Дубровина, Н.К. Прихач, В.М. Романчак // Мн.: БНТУ, 2009. – 70 с.
13. Теория вероятности и математическая статистика: учебно-методическое пособие для студентов специальностей 1-38 01 01, 1-38-01 02, 1-52 01 01, 1-38 02 02, 1-54 01 01: в 2 ч. /сост. Н.К. Прихач [и др.]; под ред. М.А. Князева. – Минск, БНТУ, 2020. – Ч. 2. – 72 с.
14. Прихач Н.К. Прикладная математика [Электронный ресурс]: учебно-методическое пособие для студентов специальности 1-54 01 01 «Метрология, стандартизация и сертификация (по направлениям)»/ Н.К. Прихач, И.В. Прусова: Белорусский национальный технический университет, кафедра инженерной математики – Минск, БНТУ, 2020 г. – Режим доступа: http://rep.bntu.by/handle/data/9383

III КОНТРОЛЬ ЗНАНИЙ

3.1. Перечень вопросов к зачету (экзамену) по дисциплине «Прикладная математика»

1. Законы распределения непрерывной случайной величины: равномерный, показательный.
2. Нормальный закон распределения непрерывной случайной величины.
3. Распределения, связанные с нормальным: χ^2, Фишера, Стьюдента.
4. Основы работы в пакете Statistica. Вероятностный калькулятор.
5. Формы, виды и способы статистического наблюдения
6. Генеральная совокупность, выборка, репрезентативность. Построение дискретного и интервального вариационного ряда.
7. Графическое изображение вариационного ряда.
8. Точечные оценки характеристик случайной величины: средние величины.
9. Структурные характеристики выборочной совокупности. Мода и медиана
10. Точечные оценки характеристик случайной величины: показатели вариации.
11. Дисперсия и среднее квадратическое отклонение
12. Точечные оценки характеристик случайной величины: асимметрия и эксцесс. Предварительная проверка на нормальность.
13. Интервальные оценки. Доверительная вероятность и доверительный интервал**.**
14. Проверка статистических гипотез. Нулевая, конкурирующая, простая и сложная гипотезы. Ошибки первого и второго рода.
15. Критерии проверки гипотез. Уровень значимости и критические области. Схема проверки статистических гипотез.
16. Приближенная проверка на нормальность (графическая, с помощью коэффициентов асимметрии и эксцесса, с использованием σ).
17. Критерий χ^2. Проверка гипотезы о нормальном распределении случайной величины. Правило Романовского.
18. Критерий Колмогорова.
19. Проверка гипотезы о значении математического ожидания нормально распределённой случайной величины.
20. Проверка гипотезы о значении дисперсии нормально распределённой случайной величины.
21. Проверка гипотез равенства математических ожиданий двух случайных величин, распределённых нормально.
22. Проверка гипотезы о дисперсиях двух случайных величин, распределённых по нормальному закону. Критерий Фишера.
23. Проверка гипотез о дисперсиях нескольких случайных величин, имеющих нормальное распределение. Критерии Бартлетта и Кохрена.

24. Непараметрическое сравнение выборочных статистик. Критерий Манна-Уитни и двухвыборочный критерий Вилкоксона.
25. Непараметрические методы математической статистики. Проверка однородности двух выборок – критерий знаков и знако-ранговый критерий Вилкоксона.
26. Функциональная, статистическая и корреляционная зависимости.
27. Линейная регрессия. Метод наименьших квадратов.
28. Остаточная дисперсия. Коэффициент детерминации. Адекватность линейной регрессии результатам наблюдений.
29. Ковариация и выборочный коэффициент корреляции. Проверка статистических гипотез о корреляционной зависимости.
30. Выборочный коэффициент ранговой корреляции Спирмена, его свойства. Проверка гипотезы о его значимости.
31. Выборочный коэффициент ранговой корреляции Кендалла, его свойства. Проверка гипотезы о его значимости.
32. Выборочное корреляционное отношение и его свойства.
33. Нелинейная регрессия. Некоторые нелинейные задачи, сводящиеся к линейным моделям.
34. Множественная регрессия. Построение линейной регрессионной модели. Коэффициент множественной корреляции, его свойства.
35. Автокорреляция остатков. Критерий Дарбина-Уотсона.
36. Дисперсионный анализ. Основные понятия.
37. Однофакторный дисперсионный анализ.
38. Двухфакторный дисперсионный анализ Основные понятия.
39. Однофакторный непараметрический анализ. Критерий Краскелла-Уоллиса.
40. Двухфакторный непараметрический анализ. Критерий Фридмана. Коэффициент конкордации (согласованности).
41. Временные ряды. Основные понятия и определения.
42. Стационарные временные ряды и их характеристики. Автокорреляционная функция.
43. Аналитическое выравнивание (сглаживание) временного ряда.
44. Прогнозирование на основе моделей временных рядов.
45. Понятие об авторегрессионных моделях и моделях скользящей средней.
46. Временной ряд, тренд, трендовая модель. Получение трендовой модели средствами Excel.

3.2. Тесты для самоконтроля знаний

3.2.1 Проверочный тест по теме «Случайные величины»

1. Величина, которая в зависимости от результата эксперимента, может принимать различные числовые значения, называется:

1) Случайной 2) Дискретной 3) Непрерывной
4) Вероятностью 5) Другой ответ

2. Случайная величина, принимающая различные значения, которые можно записать в виде конечной или бесконечной последовательности, называется:

1) Непрерывной случайной величиной 2) Постоянной величиной
3) Дискретной случайной величиной 4) Переменной величиной
5) Другой ответ

3. Дискретная случайная величина – это:

1) Случайная величина, принимающая конечное число значений;
2) Случайная величина, принимающая некоторое число значений;
3) Случайная величина, принимающая счетное число значений;
4) Случайная величина, принимающая конечное, счетное число значений;
5) Другой ответ

4. Функция распределения $F(x)$ принимает значения:

1) $[0;+\infty)$; 2) $(-\infty;+\infty)$; 3) $[0;1]$; 4) $[-1;1]$;
5) Другой ответ

5. Среди выражений: а) центр распределения; б) среднее значение; в) плотность вероятности; г) математическое ожидание – синонимами являются:

1) а), г); 2) Все, кроме а) 3) Все, кроме в)
4) б), г); 5) Другой ответ

6. Модой дискретной случайной величины называют такое значение, которое:

1) Повторяется наименьшее число раз
2) Обладает максимальной дисперсией
3) Обладает минимальной дисперсией
4) Имеет наибольшую вероятность
5) Другой ответ

7. Математическое ожидание случайной величины характеризует

1) Степень случайности
2) Наиболее вероятное распределение случайной величины
3) Степень рассеивания значений случайной величины

4) Среднее значение случайной величины
5) Другой ответ

8. Формула, по которой вычисляется математическое ожидание

1) $\sum_{i=1}^{n} x_i p_i$; 2) $M\left(X^2\right) - M^2(X)$ 3) $\sqrt{D(X)}$ 4) $\frac{N+1}{2}$

5) Другой ответ

9. Кривая нормального закона распределения называется:

1) Кривой Паскаля; 2) Кривой Гаусса; 3) Прямой Ферма;
4) Кривой Колмогорова; 5) Другой ответ

10. Дисперсия характеризует

1) Среднее значение случайной величины
2) Рассеивание, разброс случайной величины
3) Наибольшую вероятность случайной величины
4) Наименьшую вероятность случайной величины
5) Другой ответ

11. Формула, по которой вычисляется дисперсия

1) $\sum_{i=1}^{n} x_i p_i$; 2) $M\left(X^2\right) - M(X)$ 3) $M\left(X^2\right) - M^2(X)$

4) $M\left(X^2\right) + M^2(X)$ 5) Другой ответ

12. Какое распределение не относится к непрерывной случайной величине?

1) Нормальное 2) Равномерное 3) Биномиальное
4) Показательное 5) Другой ответ

13. Плотность вероятности $f(x)$ равномерно распределенной случайной величины *X* сохраняет в интервале (1; 3) постоянное значение, равное c; вне этого интервала плотность вероятности равна нулю. Найти c. В ответ записать $10 \cdot c$.

1) 4 2) 5 3) 6 4) 7 5) Другой ответ

14. Нормально распределенная случайная величина X задана плотностью $f(x) = \frac{1}{3\sqrt{\pi}} e^{-\frac{(x-1)^2}{18}}$. Дисперсия случайной величины X равна…

1) 1; 2) 9; 3) 3; 4) 5 5) Другой ответ

15. Дан ряд распределения дискретной случайной величины *X*:

x_i	–1	1	3
p_i	0,1	0,6	0,3

Дисперсия *D*(*X*) равна:
1) 1,44 2)2,16 3) 1,8
4) 0,6 5) Другой ответ

3.2.2 Проверочный тест по теме «Выборка и ее анализ»

1. Совокупность объектов, из которых производится выборка, называется….
1) Генеральной 2) Средней 3) Вероятной
4) Выборочной 5) Другой ответ

2. Для того чтобы по выборке можно было судить о случайной величине, выборка должна быть…
1) Бесповторной 2) Повторной 3) Репрезентативной
4) Безвозвратной 5) Другой ответ

3. Полигон служит для изображения…
1) Дискретного ряда 2) Гистограммы 3) Интервального ряда
4) Выборочной функции 5) Другой ответ

4. Ступенчатая фигура из прямоугольников с основаниями, равными интервалам значения признака $x_i - x_{i+1}\left(i = \overline{1,n}\right)$ и высотами, равными частотам (частостям) интервалов носит название
1) Полигона 2) Гистограммы 3) Кумулянты
4) Выборочной функции 5) Другой ответ

5. Статистическая оценка, математическое ожидание которой равно оцениваемому параметру при любом объеме выборки, называется…
1) Несмещенной 2) Состоятельной 3) Эффективной
4) Прямой 5) Обратной

6. Статистическая оценка, которая (при заданном объеме выборки) имеет наименьшую возможную дисперсию, называется
1) Несмещенной 2) Состоятельной 3) Эффективной
4) Прямой 5) Обратной

7. Интервальная оценка – это:

1) оценка параметра генеральной совокупности параметром, рассчитанным на основе выборки;

2) нахождение интервала, в который попадает наудачу брошенная точка;

3) оценка интервала вероятностей, с которыми может происходить некоторое событие;

4) оценка параметра генеральной совокупности интервалом, в который этот параметр с заданной вероятностью попадет.

5) другой ответ

8. Что является оценкой математического ожидания?

1) Выборочная средняя 2) Выборочная дисперсия

3) Исправленная дисперсия 4) Относительная частота

5) Другой ответ

9. Вариант, которому соответствует наибольшая частота вариационного ряда, называется…

1) медианой 2) модой 3) дисперсией 4) вариантом

5) другой ответ

10. Что является несмещённой оценкой генеральной дисперсии?

1) Выборочная средняя 2) Выборочная дисперсия

3) Исправленная дисперсия 4) Относительная частота

5) Другой ответ

11. Чему равна сумма доверительной вероятности и уровня значимости $\gamma+\alpha$?

1) 0 2) неотрицательному числу 3) 1

4) какому-то числу от 0 до 1 5) Другой ответ

12. 3, 3, 1, 2, 5, 4, 2, 2, 4, 0, 2, 3, 2, 0, 2 – выборка. Относительная частота варианты 2 составляет…

1) 2/5 2) 1/3 3) 1/5 4) 2/3 5) Другой ответ

13. В результате измерений некоторой физической величины одним прибором (без систематических ошибок) получены следующие результаты (в мм): 8, 11, 11. Тогда несмещенная оценка дисперсии измерений равна…

1) 2 2) 9 3) 6 4) 3 5) Другой ответ

14. Точечная оценка параметра распределения равна 21. Тогда его интервальная оценка может иметь вид…

1) (20; 21) 2) (21; 22) 3) (0; 21)

4) (20; 22) 5) Другой ответ

15. Из генеральной совокупности X с нормальным распределением извлечена выборка и составлен статистический ряд:

X	–3	–1	0	1	3	5
n	1	2	4	3	3	2

Найти доверительный интервал для математического ожидания. Принять $\gamma = 0{,}95$.

Ответ округлить до 2 цифр после запятой.

1) (–0,12; 2,38); 2) (–0,12; 2,39); 3) (–0,13; 2,38);
4) (–0,13; 2,39) 5) Другой ответ

3.2.3 Проверочный тест по теме «Проверка статистических гипотез и дисперсионный анализ»

1. Статистической гипотезой называют предположение:

1) о равенстве двух параметров 2) о неравенстве двух величин
3) о виде или параметрах неизвестного закона распределения случайной величины
4) относительно объема генеральной совокупности
5) другой ответ

2. При анализе данных выдвигаются следующие гипотезы:

1) нулевая гипотеза и гипотеза однородности
2) нулевая и альтернативная гипотезы
3) нулевая гипотеза и гипотеза равенства средних
4) гипотеза однородности и гипотеза отсутствия ошибок репрезентативности
5) другой ответ

3. Степень соответствия эмпирических и теоретических распределений вероятностей, а также двух эмпирических распределений, позволяют определить:

1) непараметрические критерии
2) параметрические и непараметрические критерии
3) параметрические критерии
4) критерии согласия
5) другой ответ

4. Область значений статистического критерия, когда нулевая гипотеза отвергается, называется:

1) критической областью; 2) полупрямой;
3) интервалом; 4) областью принятия гипотезы;
5) другой ответ

5. Вероятность статистического критерия принять верную гипотезу называется:

1) уровнем значимости; 2) уровнем доверия;
3) мощностью критерия; 4) ошибкой второго рода
5) другой ответ

6. Для корректного использования критерия Пирсона объем выборочной совокупности должен быть**:**

1) не менее 10 2) не менее 30 3) не менее 50
4) не менее 150 5) другой ответ

7. Что называют ошибкой второго рода?

1) Гипотеза H_0 верна и ее принимают согласно критерию
2) Гипотеза H_0 верна, но ее отвергают согласно критерию
3) Гипотеза H_0 неверна и ее отвергают согласно критерию
4) Гипотеза H_0 неверна, но ее принимают согласно критерию
5) Другой ответ

8. При проверке гипотезы о равенстве дисперсий двух нормально распределенных случайных величин наблюдаемое значение критерия сравнивают с критической точкой распределения:

1) Стьюдента; 2) Фишера; 3) Пирсона;
4) Гаусса; 5) нормального

9. Если принимается гипотеза $H_0 : a = a_0$ о среднем размере детали, то детали …

1) соответствуют стандарту; 2) меньше стандарта;
3) больше стандарта; 4) нельзя сделать вывод
5) другой ответ

10. На двух токарных станках обрабатываются втулки. Если принимается гипотеза о работе станков $D_1 > D_2$ (дисперсия размера втулок больше для первого станка), то…

1) первый станок налажен лучше;
2) второй станок налажен лучше
3) станки налажены одинаково;
4) нельзя сделать вывод
5) другой ответ

11. Для проверки равенства дисперсий у нескольких выборок применяется критерий:

1) Фишера; 2) Кохрена; 3) Колмогорова;
4) максимального правдоподобия; 5) Другой ответ

12. Необходимым условием для осуществления дисперсионного анализа является:

1) одинаковый размер выборок;
2) наличие более трех выборок;
3) равенство средних у всех выборок;
4) равенство дисперсий у всех выборок;
5) другой ответ

13. Какая из дисперсий выражает собой влияние неучтенных факторов на результативный признак:

1) Факторная 2) Общая 3) Остаточная
4) Выборочная 5) Другой ответ

14. Среди приведенных высказываний найдите ошибочное.

1) Дисперсионный анализ следует применять тогда, когда установлено, что распределение результативного признака является нормальным.
2) Дисперсионный анализ состоит в определении степени связи между двумя случайными величинами.
3) Дисперсионный анализ используется в тех случаях, когда есть в распоряжении три или более независимые выборки, полученные из одной генеральной совокупности
4) Все высказывания верны 5) Все высказывания неверны

15. Требуется при уровне значимости $\alpha = 0{,}05$ проверить по критерию согласия Пирсона (и по правилу Романовского) гипотезу H_0 о нормальном распределении генеральной совокупности, если известны эмпирические частоты n_i и теоретические частоты n_i':

n_i	5	12	19	27	20	10	7
n_i	6	13	13	28	21	13	6

В ответе указать значение c. Ответ округлить до 2 цифр после запятой:

1) $c = 2{,}98$; нет оснований отвергнуть гипотезу H_0;
2) $c = 3{,}00$; нет оснований отвергнуть гипотезу H_0;
3) $c = 3{,}56$; гипотеза H_0 отвергается;
4) $c = 3{,}67$; гипотеза H_0 отвергается;
5) Другой ответ

3.2.4 Проверочный тест по теме «Парный корреляционно-регрессионный анализ и нелинейная регрессия»

1. Задачей корреляционного анализа *не является*

1) установление направления корреляционной связи;
2) установление формы корреляционной связи;

3) измерение тесноты корреляционной связи;
4) нахождение уравнения регрессии.
5) другой ответ

2. Связь считается сильной, если значение выборочного коэффициента корреляции

1) Равно 0 2) В диапазоне от 0 до 0,3
3) В диапазоне от 0,7 до 1 4) В диапазоне от 1 до 2
5) Другой ответ

3. Если одному значению первого признака соответствует несколько значений второго – это связь:

1) Функциональная 2) Положительная
3) Регрессионная 4) Прямолинейная
5) Корреляционная

4. Для изображения корреляционной зависимости используется график:

1) Линейный 2) Радиальный 3) Рассеяния точек
4) Динамический 5) Другой ответ

5. Если ковариация (K_{xy}) больше нуля, то…

1) Взаимосвязь величин отсутствует;
2) Существует прямая взаимосвязь;
3) Существует обратная взаимосвязь;
4) Существует нелинейная взаимосвязь;
5) Другой ответ

6. Универсальным показателем тесноты связи между факторным и результативным признаками является:

1) коэффициент регрессии; 2) корреляционное отношение;
3) коэффициент корреляции; 4) корреляционный момент;
5) другой ответ

7. Значения корреляционного отношения η принадлежат промежутку…

1) $[0; \infty)$; 2) $[0; 1]$; 3) $[0; 2]$; 4) $[-1; 1]$; 5) Другой ответ

8. Коэффициент Спирмена является показателем связи между переменными, измеренными в шкале:

1) интервалов 2) рангов 3) наименований
4) равных отношений 5) другой ответ

9. Выборочное уравнение прямой линии регрессии *Y* на *X* имеет вид $\bar{y}_x = 1{,}4 - 1{,}8x$, а средние квадратические отклонения равны $\sigma_x = 0{,}12$; $\sigma_y = 0{,}54$. Тогда коэффициент корреляции равен…

1) –0,6 2) –0,4 3) –0,02 4) 0,4 5) Другой ответ

10. Предложенная формула $1-\frac{6\cdot\sum d_i^2}{n^3-n}$ является

1) парным коэффициентом корреляции;
2) коэффициентом корреляции рангов Спирмена;
3) коэффициентом детерминации;
4) коэффициентом корреляции рангов Кендалла
5) другой ответ

11. Какое уравнение регрессии нельзя свести к линейному виду?

1) $\bar{y}_x = a_0 + a_1 \ln x$; 2) $\bar{y}_x = a_0 + \frac{a_1}{x}$; 3) $\bar{y}_x = a_0 + a_1 x^c$;

4) $\bar{y}_x = a_0 x^{a_1}$ 5) Другой ответ

12. Уравнение регрессии связывает значения факторного признака:
1) С определенным значением результативного признака;
2) Максимальным значением результативного признака;
3) Средним значением результативного признака;
4) Дисперсией результативного признака.
5) Другой ответ

13. Согласно методу наименьших квадратов наилучшей аппроксимирующей кривой будет та, для которой:

1) Среднее отклонение ординат эмпирических точек от расчетных будет минимальным;

2) Квадрат среднего отклонения ординат эмпирических точек от расчетных будет минимальным;

3) Сумма отклонений ординат эмпирических точек от расчетных будет минимальной;

4) Сумма квадратов отклонений ординат эмпирических точек от расчетных будет минимальной.

5) Другой ответ

14. По 5 объектам получены следующие результаты:

$\sum x = 10$; $\sum x^2 = 30$; $\sum y = -52$; $\sum y \cdot x = -146{,}5$.

Уравнение регрессии Y на X имеет вид:

1) $\bar{y}_x = -4{,}25x + 1{,}9$; 2) $\bar{y}_x = 4{,}25x + 1{,}9$;

3) $\bar{y}_x = -4{,}25x - 1{,}9$; 4) $\bar{y}_x = 4{,}25x - 1{,}9$;

5) Другой ответ

15. При дегустации 10 сортов продукции двумя специалистами были получены следующие последовательности рангов:

r_i: 9 5 10 4 3 1 7 2 8 6

s_i: 8 2 7 4 3 1 10 4 6 5

Вычислить коэффициент ранговой корреляции Спирмена. Ответ округлить до двух цифр после запятой:

1) 0,78; 2) 0,75; 3) 0,73; 4) 0,80;

5) Другой ответ

3.2.5 Проверочный тест по теме «Непараметрическая статистика»

1. При проверке статистических гипотез непараметрические критерии используются

1) Только в случае, когда закон распределения значений анализируемых признаков является нормальным;

2) В случае, когда закон распределения значений анализируемых признаков неизвестен;

3) Если применение параметрических критериев не позволяет отвергнуть нулевую гипотезу;

4) Для сравнения трех и более выборок;

5) Другой ответ.

2. К непараметрическим критериям в статистике относят:

1) Критерий Стьюдента и критерий Уилкоксона;

2) Критерий Уилкоксона и критерий Манна-Уитни;

3) Критерий Фишера и критерий Манна-Уитни;

4) Критерий Стьюдента и критерий Фишера;

5) Другой ответ

3. Т-критерий Уилкоксона это…

1) Ранговый критерий для сравнения независимых выборок;

2) Ранговый критерий для сравнения зависимых выборок;

3) Параметрический критерий для сравнения независимых выборок;

4) Параметрический критерий для сравнения зависимых выборок;

5) Другой ответ

4. Критерий Манна-Уитни это….

1) Ранговый критерий для сравнения независимых выборок;

2) Ранговый критерий для сравнения зависимых выборок;

3) Параметрический критерий для сравнения независимых выборок;

4) Параметрический критерий для сравнения зависимых выборок;

5) Другой ответ

5. При использовании критерия Манна-Уитни значения эмпирическое $U = 64$, критическое $U_{kp}(\alpha = 0{,}05) = 72$ принимается гипотеза:

1) О «различии» – H_1; 2) О «сходстве» – H_0;

3) Вывод зависит от условий эксперимента;
4) С вероятностью 0,05 принимается H_0;
5) Другой ответ

6. Количество выборок, сопоставляемых в критерии Манна-Уитни:
1) Одна 2) Две 3) Три 4) Больше трех;
5) Другой ответ

7. Если гипотеза H_0 по критерию Манна–Уитни U отклоняется на уровне значимости 0,01, то на уровне значимости 0,05:
1) Отвергается; 2) Ответ зависит от гипотезы H_0
3) Принимается; 4) Ответ зависит от гипотезы H_1
5) Другой ответ

8. Для сравнения двух зависимых групп по количественному признаку вне зависимости от распределения используют
1) t-критерий Стьюдента; 2) дисперсионный анализ;
3) Т-критерий Уилкоксона; 4) тест Манна-Уитни;
5) другой ответ

9. Критерий знаков применяется для проверки нулевой гипотезы:
1) Об однородности генеральной совокупности по зависимым выборкам
2) Об однородности генеральной совокупности по независимым выборкам
3) О законе распределения генеральной совокупности
4) О значении параметров закона распределения
5) Другой ответ

10. Для сравнения двух независимых групп по количественному признаку вне зависимости от распределения используют
1) t-критерий Стьюдента; 2) Дисперсионный анализ;
3) Критерий знаков 4) Тест Манна-Уитни
5) Другой ответ

11. Для сравнения трех независимых групп по количественному вне зависимости от распределения используют
1) Критерий Уилкоксона; 2) Критерий Краскела-Уоллиса;
3) Критерий Фридмана; 4) Тест Манна-Уитни;
5) Другой ответ

12. Для сравнения трех зависимых групп по количественному вне зависимости от распределения используют
1) Критерий Уилкоксона; 2) Критерий Краскела-Уоллиса;
3) Критерий Фридмана; 4) Тест Манна-Уитни;

5) Другой ответ

13. Объем выборок, необходимых для применения критерия Краскелла-Уоллеса должен быть:

1) 4 2) ≤ 5 3) ≥ 5

4) Произвольный 5) другой ответ

14. Связь между статистиками U (Манна-Уитни) и V (Уилкоксона) определяется соотношением:

1) $V = U + \frac{n_1(n_1 - 1)}{2}$; 2) $V = U + \frac{n_1(n_1 + 1)}{2}$ 3) $V = U + \frac{n_2(n_1 - 2)}{2}$;

4) $V = U + \frac{n_1(n_1 - 1)}{2}$ 5) другой ответ

15. Кинопленка четырех видов была представлена трем экспертам для определения лучшей. Каждому эксперту предложили упорядочить пленки по степени предпочтения. Баллы (ранги), проставленные экспертами приведены в таблице. Наибольший балл соответствует пленке самого лучшего качества:

Эксперты	Вид пленки			
	1-я	2я	3-я	4-я
1-й	2	1	3	4
2-й	2	1	4	3
3-й	2	1	4	3
$\sum$	6	3	11	10

Требуется оценить, различаются ли виды пленки и согласованы ли оценки экспертов (вычислить статистику Фридмана и коэффициент конкордации)

1) $F_B = 8;\ W = 0{,}91$; 2) $F_B = 8{,}1;\ W = 0{,}9$

3) $F_B = 8{,}2;\ W = 0{,}9$; 4) $F_B = 8{,}2;\ W = 0{,}91$;

5) Другой ответ

3.3.6 Проверочный тест по теме «Задачи прогнозирования»

1. Временной ряд это:

1) Произвольно расположенные в произвольном порядке показатели, характеризующие развитие явления во времени

2) Последовательно расположенные в хронологическом порядке показатели, характеризующие развитие явления во времени

3) Последовательно расположенные по мере возрастания показатели, характеризующие развитие явления во времени

4) Последовательно расположенные по мере убывания показатели, характеризующие развитие явления во времени

5) Другой ответ

2. Вычисление групповой средней заключается в

1) Определении средней величины каждого укрупненного периода

2) Суммировании данных за ряд смежных периодов

3) Расчете средней арифметической предыдущего, данного и последующего уровней временного ряда

4) Определении общей выборочной средней

5) Другой ответ

3. Вычисление скользящей средней заключается в

1) Суммировании данных за ряд смежных периодов

2) Определении средней величины каждого укрупненного периода

3) Расчете средней арифметической предыдущего, данного и последующего уровней временного ряда

4) Определении общей выборочной средней

5) Другой ответ

4. Аддитивная модель временного ряда имеет вид:

1) $Y = T \cdot S \cdot E$ 2) $Y = T + S + E$ 3) $Y = T \cdot S + E$

4) $Y = T + S \cdot E$ 5) Другой ответ

5. Мультипликативная модель временного ряда имеет вид

1) $Y = T \cdot S \cdot E$ 2) $Y = T + S + E$ 3) $Y = T \cdot S + E$

4) $Y = T + S \cdot E$ 5) Другой ответ

6. Убывающая или возрастающая компонента временного ряда, характеризующая совокупное долговременное воздействие множества факторов, называется:

1) Трендовой компонентой; 2) Случайной компонентой

3) Циклической компонентой; 4) Сезонной компонентой;

5) Другой ответ

7. Нестационарность временного ряда y_t может проявляться:

1) Постоянством дисперсии его уровней;

2) Неизменностью функции регрессии во времени;

3) Гомоскедастичностью его остатков;

4) Наличием в его структуре тренда;

5) Другой ответ

8. Критерий Дарбина-Уотсона применяется для

1) Определения автокорреляции в остатках

2) Определения наличия сезонных колебаний

3) Для оценки существенности построенной модели

4) Нахождение остаточной дисперсии

5) Другой ответ

9. Для построения модели множественной линейной регрессии вида $\tilde{y} = a_0 + a_1 x_1 + a_2 x_2$ необходимое количество наблюдений должно быть

1) Не более 10; 2) Не менее 7; 3) Не более 14
4) Не менее 14; 5) Другой ответ

10. В каких пределах меняется множественный коэффициент корреляции R?

1) $R < -1$ 2) $-1 < R < 1$ 3) $R > 1$
4) $0 < R < 1$ 5) Другой ответ

11. При добавлении еще одной переменной в уравнение регрессии коэффициент детерминации:

1) Остается неизменным; 2) Уменьшается;
3) Не уменьшается 4) Увеличивается;
5) Другой ответ

12. Скорректированный коэффициент детерминации:

1) меньше обычного коэффициента детерминации;
2) больше обычного коэффициента детерминации;
3) равен обычному коэффициенту детерминации
4) меньше или равен обычному коэффициенту детерминации;
5) другой ответ

13. Оценка статистической значимости уравнения линейной множественной регрессии в целом осуществляется с помощью:

1) Критерия Стьюдента; 2) Критерия Фишера;
3) Критерия Дарбина-Уотсона 4) Критерия Колмогорова;
5) другой ответ

14. Оценка статистической значимости коэффициентов линейной множественной регрессии в целом осуществляется с помощью:

1) Критерия Стьюдента; 2) Критерия Фишера;
3) Критерия Дарбина-Уотсона 4) Критерия Колмогорова;
5) другой ответ

15. Пусть рассматривается зависимость прибыли предприятия от затрат на новое оборудование и технику от затрат на повышение квалификации работников. Собраны статистические данные по 6 однотипным предприятиям. Данные (в млн. ден. ед.) приводятся в таблице:

№ предприятия, i	Прибыль i-го предприятия, y_i	Затраты на новое оборудование i-го предприятия, x_{i1}	Затраты на повышение квалификации на i-м предприятии, x_{i2}
1	2	3	1
2	3	3	4
3	5	5	5
4	6	7	6
5	8	9	8
6	8	10	11

Построить двухфакторную линейную регрессию $\tilde{y} = a_0 + a_1 x_1 + a_2 x_2$. Значения коэффициентов округлить до 3 цифр после запятой.

1) $\tilde{y} = 0{,}396 - 0{,}661 x_1 + 0{,}144 x_2$;
2) $\tilde{y} = 0{,}396 - 0{,}661 x_1 - 0{,}144 x_2$;
3) $\tilde{y} = -0{,}396 + 0{,}661 x_1 + 0{,}144 x_2$;
4) $\tilde{y} = 0{,}396 + 0{,}661 x_1 + 0{,}144 x_2$;
5) Другой ответ

3.3. Ответы к тестам

№ теста	**Номер вопроса**														
	1	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**	**12**	**13**	**14**	**15**
1	1	3	4	3	4	4	4	1	2	2	3	3	2	2	1
2	1	3	1	2	1	3	4	1	2	3	4	1	1	4	2
3	3	2	4	1	2	2	4	2	4	4	2	4	3	2	4
4	4	3	3	3	2	2	2	2	2	2	5	3	4	3	1
5	2	2	2	1	3	2	1	3	5	4	2	3	3	2	4
6	2	1	3	2	1	1	4	1	5	4	3	1	2	1	4

IV ВСПОМОГАТЕЛЬНЫЙ РАЗДЕЛ

4.1. Учебная программа для учреждения высшего образования по учебной дисциплине «Прикладная математика» для специальности 6-05-0716-01

Белорусский национальный технический университет

УТВЕРЖДАЮ
Проректор по учебной работе
Белорусского национального
технического университета
Ю.А. Николайчик
08.07. 2024
Регистрационный № УД-ЕСТ85-5 /уч.

ПРИКЛАДНАЯ МАТЕМАТИКА

Учебная программа учреждения высшего образования
по учебной дисциплине для специальности
6-05-0716-01 «Метрология, стандартизация и контроль качества»

Минск 2023г.

Таблица 4.1

Очная (дневная) форма получения высшего образования					
Курс	Семестр	Лекции, ч.	Лабораторные занятия, ч.	Практические занятия, ч.	Форма текущей аттестации
2	3	34	34		Зачет

Учебно-методическая карта учебной дисциплины
Очная (дневная) форма получения высшего образования
для специальности 6-05-0716-01

Номер раздела, темы	Название раздела, темы	Количество аудиторных часов					Количество часов УСР	Форма контроля знаний
		Лекции	Практические занятия	Семинарские занятия	Лабораторные занятия	Иное		
1	2	3	4	5	6	7	8	9
	3 семестр							
1.	Случайные величины							
1.1	Основные понятия теории вероятностей. Случайные величины и их числовые характеристики.	2						
	Лабораторная работа № 1. Числовые характеристики случайных величин. Вероятностные распределения. Знакомство с пакетом Statistica.				2			Защита лабораторной работы
2.	Выборка и ее анализ.							
2.1	Статистические оценки параметров распределения.	4						
	Лабораторная работа № 2. Первичная обработка статистических данных. Точечные и интервальные оценки параметров распределения				4			Защита лабораторной работы
3.	Проверка статистических гипотез и дисперсионный анализ.							
3.1	Статистическая проверка истинности выдвинутой гипотезы.	4						
	Лабораторная работа № 3. Статистическая проверка непараметрических гипотез. Критерии согласия.				4			Защита лабораторной работы

3.2	Проверка параметрических гипотез.	4						
	Лабораторная работа № 4. Проверка гипотез о параметрах распределения				4			Защита лабораторной работы
3.3	Дисперсионный анализ	2						
	Лабораторная работа № 5. Дисперсионный анализ.				2			Защита лабораторной работы
4.	Парный корреляционно-регрессионный анализ и нелинейная регрессия.							
4.1	Корреляционный анализ.	4						
	Лабораторная работа № 6. Корреляционный анализ.				4			Защита лабораторной работы
4.2	Регрессионный анализ	6						
	Лабораторная работа № 7. Регрессионный анализ				6			Защита лабораторной работы
5.	Непараметрическая статистика							
5.1	Непараметрические методы математической статистики	4						
	Лабораторная работа № 8.Непараметрические методы математической статистики				4			Защита лабораторной работы
6	Задачи прогнозирования							
6.1	Временные ряды и множественная линейная регрессия	4						
	Лабораторная работа № 9. Анализ и прогнозирование временных рядов. Множественная линейная регрессия.				4			Защита лабораторной работы
	Итого за семестр	34			34			зачет
	Всего аудиторных часов	68						