Министерство образования Республики Беларусь
БЕЛОРУССКИЙ НАЦИОНАЛЬНЫЙ ТЕХНИЧЕСКИЙ УНИВЕРСИТЕТ

Кафедра «Инженерная математика»

ПРИКЛАДНАЯ МАТЕМАТИКА

Учебно-методическое пособие для студентов специальности 1-54 01 01
«Метрология, стандартизация и сертификация (по направлениям)»

Электронный учебный материал

Минск БНТУ 2020

УДК 519.25 (076.5)

ББК 22.172 я 7

П 77

Авторы
Прихач Н.К.
Прусова И.В.

Рецензент
Щербакова Е.Н.

Белорусский национальный технический университет
пр-т Независимости, 65, г. Минск, Республика Беларусь
Тел.(017) 292-67-84
E-mail: nkprikhach@mail.ru
http://www.bntu.by/ru/struktura/facu1t/psf/im/
Регистрационный № БНТУ/ПСФ85 – 27.2020

Учебное пособие содержит восемь лабораторных работ. Рассматриваются примеры автоматизации статистических вычислений с помощью программных продуктов Excel и STATISTICA. Приводятся типовые задачи с использованием компьютерных технологий, а также перечень вариантов заданий для самостоятельного выполнения. Учебное пособие может быть использовано для самостоятельной работы студентами как заочного, так и дневного отделения специальности, изучающих дисциплину «Прикладная математика».

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Многие современные исследования базируются на методах математической статистики. Развитие компьютерных информационных технологий позволяет поднять использование этих методов на новый уровень.

Умение правильно выбрать метод решения задачи, учесть его возможности и ограничения, грамотно применить и интерпретировать результат – основные требования к специалисту, который использует статистические методы. Поэтому дисциплина «Прикладная математика» имеет две главные цели: познакомить студентов с теоретическими основами статистической обработки данных и научить их автоматизировать этот процесс с помощью компьютера.

Пособие рассчитано на 68 часов аудиторных занятий (в том числе 34 часа лабораторных). Излагаются основные понятия, приемы и методы статистической обработки данных наблюдений. Теоретический материал сопровождается большим количеством примеров. В качестве основного инструмента статистического анализа используются возможности пакетов Excel и STATISTICA.

По каждой теме даются типовые задачи для лабораторных работ, предусмотренных учебной программой. Темы лабораторных работ соответствуют основным разделам учебной программы.

В первой работе приводится описание основных характеристик обрабатываемых выборок и основных понятий математической статистики. Примеры прикладных исследований реализуются в конкретных пакетах анализа.

Далее даются расширенные определения таких понятий как нулевая и альтернативная гипотезы и объясняется суть статистической гипотезы. Рассматриваются особенности процесса проверки статистических гипотез и даются определения таким терминам, как: степени свободы, уровень значимости, мощность критериев. Дается описание и примеры проверки гипотез о законах распределения значений исследуемой выборки. Ведется оценка равенства средних и дисперсий. Для каждого из критериев даны разъяснения о назначении критерия и порядке его расчета.

Также рассматривается вопрос об оценке влияния факторов на исследуемый признак. Материал содержит подробные примеры для зависимых и независимых выборок. Дается подробное описание процесса подготовки данных к дисперсионному анализу с привлечением компьютерных пакетов.

Затем представлен материал по исследованию взаимосвязи между изучаемыми признаками. Вводится понятие коэффициента корреляции Пирсона. Также приводится описание, как параметрического коэффициента корреляции, так и непараметрических мер связи: ранговых коэффициентов корреляции Спирмена и Кендалла.

И в заключении приведены основные подходы к прогнозированию на основе анализа временных рядов.

Лабораторная работа № 1

ПЕРВИЧНАЯ ОБРАБОТКА СТАТИСТИЧЕСКИХ ДАННЫХ

Цель работы: привить навыки первичной обработки эмпирических данных с помощью методов математической статистики: получение из выборочных данных эмпирического закона распределения исследуемых признаков, событий, процессов; вычисление числовых характеристик этих распределений.

Используемые программные средства: MS Excel 2010 (2016), STATISTICA 8.0.

1.1. Краткие теоретические сведения

Вариационные ряды и их графическое изображение.

Любое статистическое исследование начинается со сбора данных об исследуемом случайном объекте. Таким объектом может быть случайное событие, случайная величина, система случайных величин (случайный вектор) или случайная функция. Этот этап работы называют *наблюдением*. Данные, собранные и зафиксированные в ходе наблюдения, называются *данными наблюдения*.

Наиболее часто объектом исследования является какая-либо случайная величина: время прохождения сообщения от отправителя до адресата, время безотказной работы технического устройства, число знаков (символов) в сообщении и т. п.

Пусть в одинаковых условиях и независимо друг от друга производится n измерений случайной величины X. Пусть $x_1, x_2, \dots, x_n$ – результаты измерений, которые называются *наблюденными значениями* или *реализациями* этой случайной величины. Наблюденные значения исследуемой случайной величины в статистике принято рассматривать как *случайную выборку* из бесконечной *генеральной совокупности* реализаций этой случайной величины, которые могли быть получены при проведении всех наблюдений над этой случайной величиной. Выборка является основным исходным объектом любого статистического исследования.

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

Наблюдаемое значение x_i называют *вариантой*, а их последовательность, записанную в возрастающем порядке, – *вариационным рядом.*

Основное назначение математико-статистических методов о состоит в том, чтобы с их помощью на основании ограниченного числа выборочных данных получить как можно более полное представление об изучаемых случайных величинах. Например, основываясь на анализе выборки, полученной в ходе наблюдения над случайной величиной X, закон распределения которой

неизвестен, сделать обоснованное заключение о функции распределения $F\ x$ этой случайной величины или об её числовых характеристиках.

Для того чтобы по имеющейся выборке можно было сделать обоснованный вывод о свойствах всей генеральной совокупности, она должна быть *репрезентативной (представительной)*, т. е. хорошо отображать свойства исследуемой генеральной совокупности. Репрезентативность выборки достигается отсутствием всякой предвзятости (вольной или невольной) по отношению к отбираемым элементам. Каждый элемент генеральной совокупности должен иметь равную со всеми элементами возможность включения их в выборку.

Если изучается *дискретная случайная величина*, число различных наблюденных значений которой не велико, то для каждого из отличающихся друг от друга значений x_i подсчитываются частоты n_i и относительные частоты (частости) $\frac{n_i}{n}$ появления этих значений в выборке.

Результаты вычислений заносятся в таблицу 1.1, которая называется *сгруппированным статистическим рядом.*

Таблица 1.1 – Сгруппированный статистический ряд

x_i – наблюденные значения	x_1	x_2	…	x_k	$k \le n$
n_i – частоты	n_1	n_2	…	n_k	$\sum_{i=1}^{k} n_i = n$
$p_i^* = \frac{n_i}{n}$ – относительные частоты	$\frac{n_1}{n}$	$\frac{n_2}{n}$	…	$\frac{n_k}{n}$	$\sum_{i=1}^{k} \frac{n_i}{n} = 1$

Если изучается непрерывная случайная величина либо дискретная случайная величина, число различных значений которой достаточно велико, то интервал всех наблюденных значений разбивается на l интервалов длины h, и подсчитывается число вариант, попавших в каждый из интервалов. Результаты расчетов заносятся в таблицу 1.2, которая называется *интервальным статистическим рядом*.

Таблица 1.2 – Интервальный статистический ряд

Границы	$x_1;\ x_2$	$x_2;\ x_3$	…	$x_l;\ x_{l+1}$	
Среднее значение интервала x_i^*	x_1^*	x_2^*	…	x_l^*	
n_i – частоты	n_1	n_2	…	n_l	$\sum_{i=1}^{l} n_i = n$
$p_i^* = \frac{n_i}{n}$	$\frac{n_1}{n}$	$\frac{n_2}{n}$	…	$\frac{n_l}{n}$	$\sum_{i=1}^{l} \frac{n_i}{n} = 1$

Длину интервала – h – проще выбирать одинаковой. Практика показывает, что число интервалов рационально выбирать порядка 5-15. Для нахождения длины интервала можно воспользоваться *формулой Стерджеса:*

$$h = \frac{x_{max} - x_{min}}{1+3{,}322 \cdot \lg n} = \frac{R}{l} \qquad (1.1)$$

Если в результате вычисления по формуле (1.1) длина интервала получится дробным числом, то выбирают либо близкое целое число, либо близкую простую дробь.

Статистический ряд представляет собой первичную форму записи статистического материала и может быть обработан различными способами. Одним из способов такой обработки является построение эмпирической функции распределения случайной величины. Обозначим через n_x число наблюдений, при которых значения вариант оказываются меньше, чем x.

Эмпирической (*статистической*) *функцией распределения* называется функция $F^*\ x$, определяющая для каждого значения x относительную частоту события $X < x$:

$$F^*\ x\ = \frac{n_x}{n} = p^*\ X < x \qquad (1.2)$$

Для того чтобы найти значение эмпирической функции распределения при данном x достаточно подсчитать число опытов, в которых величина X приняла значение меньше, чем x и разделить на общее число произведенных опытов n.

Важнейшее свойство эмпирической функции распределения состоит в том, что при увеличении объема выборки n значение этой функции в каждой точке приближается к значению функции распределения $F\ x$ в указанной точке, т. е. эмпирическая функция распределения является экспериментальным аналогом (оценкой) неизвестной функции распределения.

Кумулятивная кривая (кумулянта) – ломаная, соединяющая точки с координатами $x_i;\ n_{x_i}$ или $x_i;\ \frac{n_{x_i}}{n}$, 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.

Графически статистический ряд можно представить в виде полигона частот или относительных частот. *Полигоном частот или относительных частот* называют ломаную линию, отрезки которой соединяют точки $x_i; n_i$

(соответственно $x_i;\ p_i^*$). Полигоны обычно служат для изображения выборки в случае дискретных случайных величин (рис. 1.1):

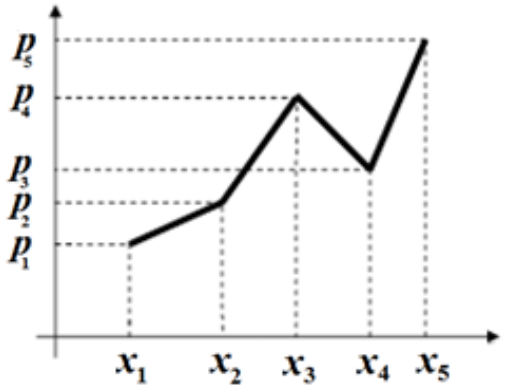


Рис. 1.1 – Полигон относительных частот

Интервальный статистический ряд часто оформляется графически в виде гистограммы. *Гистограммой* называется ступенчатая фигура (рис.1.2), состоящая из прямоугольников, основаниями которых служат отрезки, равные длине интервала, а высотами являются относительные частоты, поделенные на длину интервала. Гистограмма обычно служат для изображения выборки в случае непрерывных случайных величин. Площадь гистограммы равна единице. Если на гистограмме соединить середины верхних сторон прямоугольников, то полученная ломаная образует полигон относительных частот.

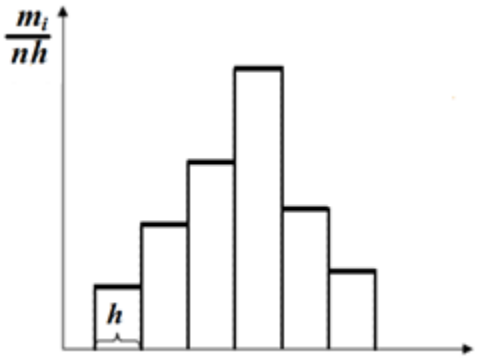


Рис. 1.2 – Гистограмма относительных частот

Точечные оценки характеристик случайной величины.

По выборочным данным часто требуется определить (приближённо найти) параметры распределения исследуемой случайной величины. Выборочная оценка θ некоторого параметра θ (например, математического ожидания или дисперсии) сама является случайной величиной и должна удовлетворять определённым требованиям: быть несмещённой, состоятельной и эффективной.

Оценка θ называется *несмещенной* (оценкой без систематической ошибки), если ее математическое ожидание при любом n равно оцениваемому па-

раметру: $M\ \theta\ = \theta$.

Оценка θ называется *состоятельной*, если при неограниченном увеличении выборки она сходится по вероятности к оцениваемому параметру: $\lim_{n\to\infty} P\ \ \theta - \theta\ < \varepsilon\ = 1$ для любого $\varepsilon > 0$.

Оценка называется *эффективной* (в некотором классе оценок), если она имеет минимальную дисперсию в этом классе.

Для вычисления точечных статистических оценок справедливы следующие формулы:

1) *Выборочная средняя*

$$x = \frac{1}{n}\ \ \sum_{i=1}^{l} x_i\, n_i = \ \ \sum_{i=1}^{l} x_i\, p_i^*$$

Здесь l – число групп для дискретного вариационного ряда (число интервалов – для интервального).

Выборочная средняя является несмещенной, состоятельной и асимптотически эффективной $\lim_{n\to\infty} D\ \theta\ = 0$ оценкой математического ожидания генеральной совокупности.

2) *Выборочная дисперсия*

$$D_B = \frac{1}{n}\ \ \sum_{i=1}^{l} x_i - x\ ^2\, n_i = \ \ \sum_{i=1}^{l} x_i - x\ ^2\, p_i^* = \ \ \sum_{i=1}^{l} x_i^2 p_i^* - x^2$$

Выборочная дисперсия является состоятельной, но смещенной оценкой дисперсии генеральной совокупности. Несмещенной оценкой дисперсии является *исправленная выборочная дисперсия*.

Исправленная выборочная дисперсия s^2 и исправленное выборочное среднее квадратическое отклонение s вычисляются по формулам:

$$s^2 = \frac{n}{n-1} D_B; \quad s = \ \ \overline{s^2}$$

3) *Стандартная ошибка среднего* оценивает изменчивость выборочного среднего, приближённо показывая, насколько выборочное среднее отличается от среднего генеральной совокупности:

$$s_x = \frac{s}{\overline{n}}$$

4) *Медианой* m_e называется вариант, который приходится на середину ряда распределения:

$$m_e = \begin{cases} x_{k+1}, \text{если } n = 2k+1 \\ \dfrac{x_k + x_{k+1}}{2}, \text{если } n = 2k \end{cases}$$

При вычислении медианы интервального ряда распределения используется формула:

$$m_e = x_0 + h \cdot \frac{\frac{1}{2}n - S_{j-1}}{n_j}$$

где x_0 – начальное значение интервала, который содержит медиану, S_{j-1} – накопленная частота интервала, предшествующего медианному (сумма частот), включая интервал, предшествующий медианному, n_j – частота медианного интервала; h – величина медианного интервала.

Медианный интервал – такой интервал, которому соответствует кумулятивная (накопленная) частота, равная или впервые превышающая половину суммы частот.

5) *Модой* M_o вариационного ряда называется варианта, которой соответствует наибольшая частота. Если распределение интервальное, то мода вычисляется по формуле:

$$M_o = x_0 + h \cdot \frac{n_i - n_{i-1}}{n_i - n_{i-1} \; + \; n_i - n_{i+1}}$$

где x_0 – начальное значение модального интервала, т.е. интервала, который содержит моду; h – величина модального интервала; n_{i-1} – частота интервала, предшествующего модальному; n_{i+1} – частота интервала, следующего за модальным; n_i – частота модального интервала.

В непрерывном распределении с равными интервалами модальным будет тот интервал, которому соответствует наибольшая частота.

Если интервалы неравные, то модальным будет интервал, у которого наибольшая плотность распределения.

6) Скошенность кривой называется *асимметрией.* Для выборочной асимметрии A_s справедлива формула

$$A_s = \frac{{}_i n_i \; x_i - x \; ^3}{ns^3} = \frac{n}{n-1 \quad n-2} \quad {}_i \quad \frac{x_i - x}{s} \; ^3$$

7) Отклонение крутизны называют *эксцессом*. Выборочный эксцесс E_x определяется формулой:

$$E_x = \frac{\sum_i n_i\ (x_i - x)^4}{ns^4} - 3 = \frac{n\ (n+1)}{(n-1)\ (n-2)\ (n-3)} \sum_i \frac{(x_i - x)^4}{s} - \frac{3\ (n-1)^2}{(n-2)\ (n-3)}$$

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*

$$V = \frac{s}{x} \cdot 100\%$$

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.

9) $R = x_{max} - x_{min}$ – *размах варьирования.*

1.2. Практическая часть

Контрольный пример. По извлеченной случайной выборке генеральной непрерывной случайной величины X:

13,6	5,9	8,2	9,4	3,5	5,1	10,2	16,5	13,8	16,3
10,3	8,2	12,8	8,3	2,2	15,7	1,3	12,6	18,1	21,5
10,9	10,7	11,9	5,2	12,2	17,9	10	6,4	13	10,4
6	20,8	9,1	13,1	14,2	7,4	13,4	4,2	5,7	12,6
12,2	6,5	6,8	15,2	15,4	16,7	9,8	7,9	9,6	13,4

- составить группированный (интервальный) ряд распределения;
- построить эмпирическую функцию распределения, ее график и кумулянту;
- вычислить эмпирические плотности распределения, построить гистограмму и полигон;
- получить точечные статистические оценки параметров распределения;

- выдвинуть гипотезу о законе генерального распределения.

Решение.

1. Располагаем значения результатов эксперимента в порядке возрастания, т.е. записываем вариационный ряд:

1,3	2,2	3,5	4,2	5,1	5,2	5,7	5,9	6	6,4	6,5	6,8	7,4	7,9	8,2
8,2	8,3	9,1	9,4	9,6	9,8	10	10,2	10,3	10,4	10,7	10,9	11,9	12,2	12,2
12,6	12,6	12,8	13	13,1	13,4	13,4	13,6	13,8	14,2	15,2	15,4	15,7	16,3	16,5
16,7	17,9	18,1	20,8	21,5										

Находим размах варьирования $R = 21{,}5 - 1{,}3 = 20{,}2$. По условию объём выборки $n = 50$. Определим оптимальную длину частичного интервала:

$$h = \frac{R}{l} = \frac{R}{1 + 3{,}322 \cdot \lg n} = \frac{20{,}2}{1 + 3{,}322 \cdot \lg 50} \approx \frac{20{,}2}{7} \approx 2{,}9$$

Тогда

$x_0 = 1{,}3;\ x_1 = 1{,}3 + 2{,}886 = 4{,}186;\ \ x_2 = 4{,}186 + 2{,}886 = 7{,}072;$
$x_3 = 9{,}958;\ x_4 = 12{,}844;\ x_5 = 15{,}73;\ x_6 = 18{,}616;\ x_7 = 21{,}502.$

Вариационный ряд представим в виде следующей таблицы:

1	**2**	**3**	**4**	**5**	**6**	**7**
1,3 – 4,186	**4,186 – 7,072**	**7,072 – 9,958**	**9,958 – 12,844**	**12,844 – 15,73**	**15,73 – 18,616**	**18,616– 21,502**
1,3	4,2	7,4	10	13	16,3	20,8
2,2	5,1	7,9	10,2	13,1	16,5	21,5
3,5	5,2	8,2	10,3	13,4	16,7	
	5,7	8,2	10,4	13,4	17,9	
	5,9	8,3	10,7	13,6	18,1	
	6	9,1	10,9	13,8		
	6,4	9,4	11,9	14,2		
	6,5	9,6	12,2	15,2		
	6,8	9,8	12,2	15,4		

12,6
12,6

Тогда интервальный статистический ряд:

Границы	1	2	3	4	5	6	7	**Σ**
x_i^*	2,743	5,629	8,515	11,401	14,287	17,173	20,059	
n_i	3	9	9	12	10	5	2	**50**
p_i^*	0,06	0,18	0,18	0,24	0,20	0,10	0,04	**1**

2. *Вычисления в пакете STATISTICA.* После запуска пакета на экране появится сетка-таблица, которую преобразуем к размерам 1×50, выполнив следующие действия:

1) нажав кнопку $Vars$ (на экране), в раскрывающемся меню выберем *Delete*; появится окно $Delete\ Variables$.
2) укажем, какие переменные-столбцы убрать;
3) нажимаем кнопку *Cases*, выбираем опцию *Add* (добавление), появится окно *Add Cases*: укажем, сколько строк добавить и куда.

Далее нужно выделить столбец – переменную $Var1$ (щелчком мыши по её заглавию) – нажать правую клавишу – в открывшемся меню выбрать *Variable Specs* (спецификации переменной) – в появившемся окне *Variable 1* ввести *Name* X.

Зададим исходные данные (на рис. 1.3 указана часть массива исходных данных).

	1 X
1	13,6
2	10,3
3	10,9
4	6
5	12,2
6	5,9
7	8,2
8	10,7
9	20,8
10	6,5
11	8,2

Рис. 1.3 – Часть массива исходных данных

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 квадратическое 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.

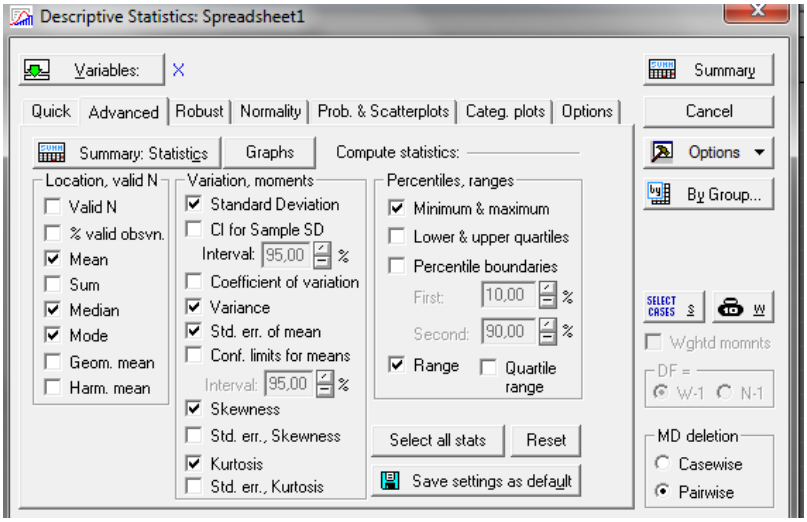


Рис. 1.4 – Вкладка *Advanced* окна *Descriptive Statistic*

Результаты вычислений размещаются в активном окне внизу (при анализе данных столбца) или слева (при анализе данных строки) от исходных данных (рис. 1.5).

	Descriptive Statistics (Spreadsheet1)									
Variable	**Mean**	Median	Mode	Frequency of Mode	Range	Variance	Std.Dev.	Standard Error	Skewness	Kurtosis
X	10,84200	10,55000	Multiple	2	20,20000	21,39024	4,624958	0,654068	0,134007	-0,323191

Рис. 1.5 – Вычисленные параметры описательной статистики

Далее, в окне *Descriptive Statistic* во вкладке *Quick* нажмём на кнопку *Frequency Table*. В результате получим таблицу частот (рис. 1.6). В первом столбце заданы интервалы для переменной X, причём последняя строка содержит пропущенные значения. Второй столбец содержит число попаданий переменной в интервалы (*Count*), третий столбец – накопленную частоту (*Cumulative Count*), четвёртый 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*).

Category	Count	Cumulative Count	Percent of Valid	Cumul % of Valid	% of all Cases	Cumulative % of All
Frequency table: X (Spreadsheet1) K-S d=,05548, p> .20; Lilliefors p> .20						
-5,00000<x<=0,000000	0	0	0,00000	0,0000	0,00000	0,0000
0,000000<x<=5,000000	4	4	8,00000	8,0000	8,00000	8,0000
5,000000<x<=10,00000	18	22	36,00000	44,0000	36,00000	44,0000
10,00000<x<=15,00000	18	40	36,00000	80,0000	36,00000	80,0000
15,00000<x<=20,00000	8	48	16,00000	96,0000	16,00000	96,0000
20,00000<x<=25,00000	2	50	4,00000	100,0000	4,00000	100,0000
Missing	0	50	0,00000		0,00000	100,0000

Рис. 1.6 – Таблица частот

Для построения графика частот (полигона частот) выделим столбец *Percent of valid*, нажмём 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. 1.7.

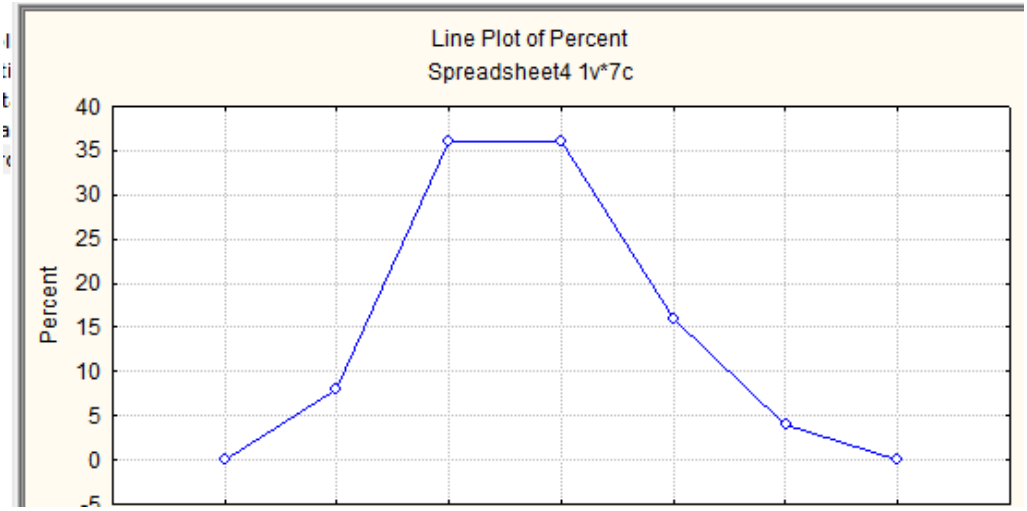


Рис. 1.7. График, построенный по исходным данным

Для построения гистограммы частот перейдём в окно *Graphs* на вкладку *Histograms*. Во вкладке *Advanced* выберем *Fit type – Normal,* установим: *Graph type: Regular, Showing type: Standard, Variables* – X*; Categories (число интервалов группировки) – 7 – OK.* Получим гистограмму, представленную на рис. 1.8.

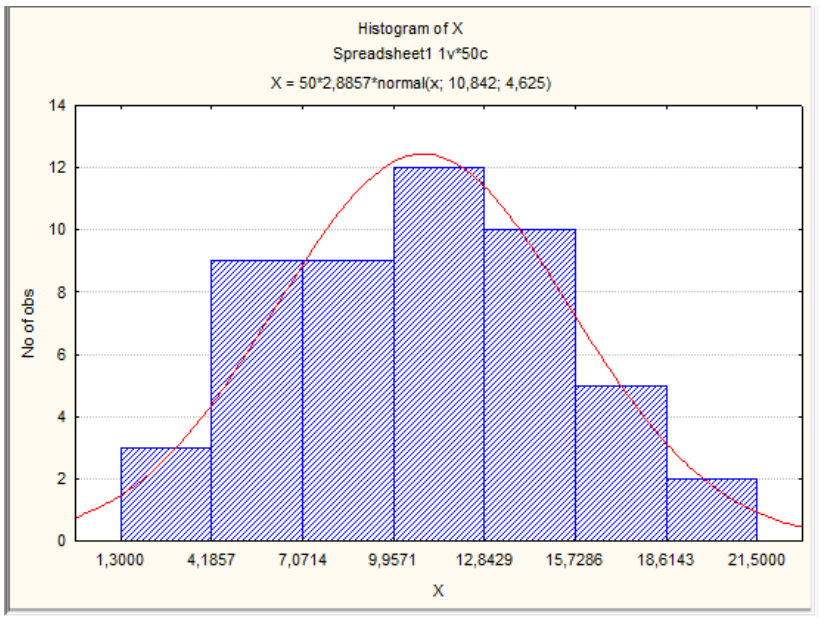


Рис.1.8 – Гистограмма, построенная по исходным данным

Для построения кумулятивной кривой во вкладке *Advanced* выберем *Fit type – Off* , установим: *Graph type: Regular, Showing type: Cumulative, Variables* – X*; Categories (число интервалов группировки) – 250 – OK.*

Наблюдаем график кумулятивной кривой (рис. 1.9).

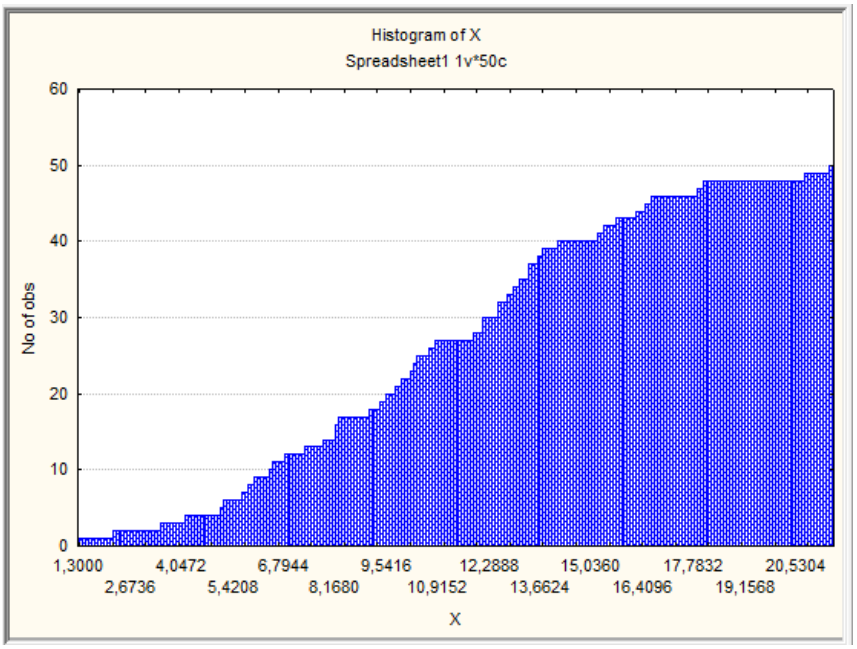


Рис. 1.9 – Кумулятивная кривая

3. *В пакете Excel* в ячейку A1 введем слово «Наблюдения» (рис. 1.10), а в диапазон A2: A51 – исходные данные.

Рассчитаем максимальное и минимальное значения выборочных данных в ячейках D1 и D2, введя соответственно функции MAX(A2: A51) и МИН(A2: A51) (рис. 1.10).

Для построения интервального вариационного ряда разобьём диапазон наблюдавшихся значений $1.3; 21.5$ на интервалы шириной $2{,}886$ (см. подпункт 1 данного примера). При этом минимальное значение должно попасть внутри интервала. В ячейку E1 введём заголовок «Варианты», а ниже, в столбце – правые границы интервалов.

В ячейке F1 запишем заголовок «Абсолютные частоты». В этом столбце будут рассчитаны значения частоты попадания в интервал. Для заполнения столбца абсолютных частот можно использовать стандартную функцию ЧАСТОТА(). 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() даёт количество наблюдений, меньших или равных ему.

Выделим мышью диапазон F2: F8, 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: A51;
- *Массив интервалов* – это диапазон значений вариант E2: E8.

Закончить ввод функции нужно одновременным нажатием клавиш $Ctrl + Shift + Enter$, поскольку ее результатом является диапазон значений.

В ячейке F9 найдём объём выборки, просуммировав значения в столбце абсолютных частот (см. рис. 1.10).

В ячейке G1 запишем заголовок «Относительные частоты». Для расчёта относительных частот внесём в ячейку G2 формулу $= F2/\$F\$9$ 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 введём формулу $H2 + G3$. Методом автозаполнения скопируем введённую формулу вниз по столбцу в диапазон H4: H8.

Итоговый вид таблицы после форматирования показан на рис. 1.10.

F2 | {=ЧАСТОТА(A2:A51;E2:E8)}

Книга1.xlsx

	A	B	C	D	E	F	G	H
1	Наблюдения		Максимум	21,5	Варианты	Абсолютные частоты	Относительные частоты	Накопленные частоты
2	13,6		Минимум	1,3	4,186	3	0,06	0,06
3	10,3				7,072	9	0,18	0,24
4	10,9				9,958	9	0,18	0,42
5	6				12,844	12	0,24	0,66
6	12,2				15,73	10	0,2	0,86
7	5,9				18,616	5	0,1	0,96
8	8,2				21,502	2	0,04	1
9	10,7			Всего наблюдений		50	1	

Рис. 1.10 – Результаты расчёта частот для интервального ряда

Построим полигон частот по данным в столбце «*Абсолютные частоты*», как показано на рис. 1.11 (используем диаграмму типа «*Точечная с прямыми отрезками и маркерами*»).

Рис. 1.11 – Полигон частот

Для построения интервального вариационного ряда можно также использовать процедуру *Гистограмма* надстройки *Пакет анализа*.

Скопируем на чистый лист данные наблюдений из столбца А, а также диапазон значений вариант из столбца Е (например, в столбец С, см. рис. 1.13).

Зададим команду *Данные/Анализ данных* и выберем инструмент анализа *Гистограмма*. Заполним окно *Гистограмма*, как показано на рис. 1.12.

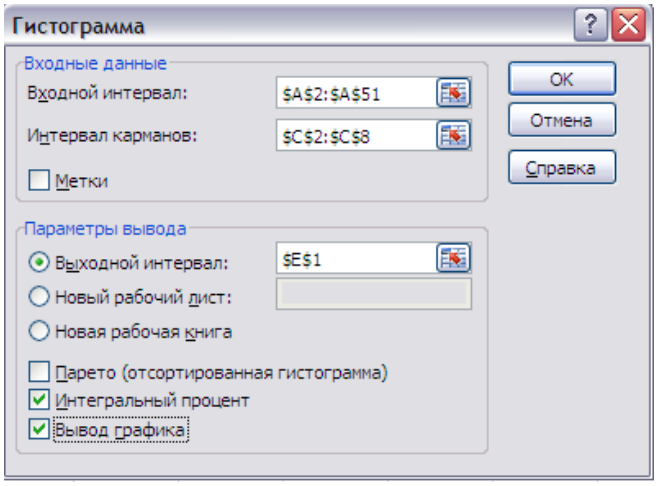


Рис. 1.12 – Диалоговое окно для построения гистограммы.

В поле *Входной интервал* укажем диапазон исследуемых данных наблюдений (A2:A51).

В поле *Интервал карманов* зададим диапазон граничных значений, опре-

деляющих выбранные интервалы (карманы). В нашем случае это диапазон возможных значений вариант (C2:C8). Они должны быть введены в возрастающем порядке. Процедура *Гистограмма* вычислит число попаданий данных между началом интервала и соседним большим по порядку. При этом включаются значения на нижней границе интервала и не включаются на верхней.

Переключатель *Параметры вывода* установим в положении *Выходной интервал* и в соответствующем поле укажем адрес ячейки, в которую будет помещён левый верхний угол результирующей таблицы (E1).

Следует также установить флажки *Вывод графика* и *Интегральный процент* (для дополнительного вывода накопленных частот).

После нажатия кнопки *OK* на рабочем листе Excel появляются таблица и диаграмма (рис. 1.13). В столбце «*Интегральный процент*» показаны накопленные частоты в процентном формате.

	A	B	C	D	E	F	G	H
1	**Наблюдения**		**Варианты**		*Карман*	*Частота*	*Интегральный %*	
2	13,6		**4,186**		4,186	3	6,00%	
3	10,3		**7,072**		7,072	9	24,00%	
4	10,9		**9,958**		9,958	9	42,00%	
5	6		**12,844**		12,844	12	66,00%	
6	12,2		**15,73**		15,73	10	86,00%	
7	5,9		**18,616**		18,616	5	96,00%	
8	8,2		**21,502**		21,502	2	100,00%	
9	10,7				Еще	0	100,00%	
10	20,8							
11	6,5							
12	8,2							
13	12,8							
14	11,9							
15	9,1							
16	6,8							
17	9,4							
18	8,3							
19	5,2							
20	13,1							
21	15,2							

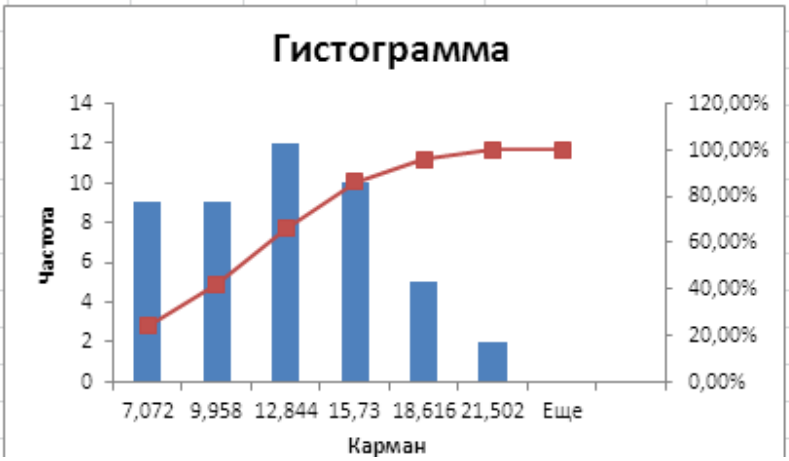


Рис. 1.13 – Таблица и диаграмма, созданные надстройкой *Пакет анализа*

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;

СТАНДОТКЛОН.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 даёт возможность получить все выборочные характеристики случайной величины.

Введём эмпирические данные на чистый лист Excel и оформим его, как показано на рис. 1.14.

	A	B	C	D	E	F	G	H
1	Наблюдения		Выборочные оценки (используя стандартные функции)				Выборочные оценки (пакет анализа)	
2	13,6		Среднее	10,842			*Наблюдения*	
3	10,3		Выборочная дисперсия	20,962				
4	10,9		Исправленная дисперсия	21,390			Среднее	10,842
5	6		Стандартное отклонение	4,625			Стандартная ошибка	0,654
6	12,2		Мода	12,200			Медиана	10,550
7	5,9		Медиана	10,550			Мода	12,200
8	8,2		Эксцесс	-0,323			Стандартное отклонение	4,625
9	10,7		Асимметрия	0,134			Дисперсия выборки	21,390
10	20,8		Количество наблюдений	50			Эксцесс	-0,323
11	6,5						Асимметричность	0,134
12	8,2						Интервал	20,200
13	12,8						Минимум	1,300
14	11,9						Максимум	21,500
15	9,1						Сумма	542,100
16	6,8						Счет	50

Рис.1.14 – Лист Excel с расчётом выборочных характеристик распределения

В ячейки D2:D9 введём стандартные функции Excel категории *Статистические:*

СРЗНАЧ, ДИСП.Г, ДИСП.В, СТАНДОТКЛОН.В, МОДА.ОДН, МЕДИАНА, ЭКСЦЕСС, СКОС и СЧЕТ.

Аргументами всех этих функций является диапазон выбранных значений A2:A51.

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. 1.15.

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»).

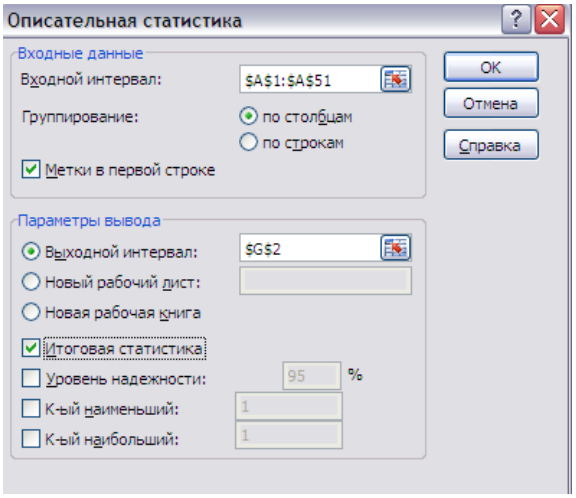


Рис. 1.15 – Диалоговое окно для вывода описательной статистики

Переключатель *Параметры вывода* установлен в положение *Выходной интервал*, так как нужно получить результаты расчётов 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 *ОК*. Результаты расчетов с помощью *Пакета анализа* показаны на рис. 1.22 в столбцах G и H.

Предположение о характере генерального распределения

Так как для нормального распределения эксцесс и асимметрия равны нулю, то достаточно малые значения E_x и A_s, вид гистограммы и полигона, а также то, что мода и медиана близки к среднему выборочному, позволяют выдвинуть гипотезу о нормальном распределении генеральной совокупности.

1.3. Задания для самостоятельной работы.

На основе совокупности опытных данных выполнить следующие задания:

- составить группированный (интервальный) ряд распределения;
- построить эмпирическую функцию распределения, ее график и кумулянту;
- вычислить эмпирические плотности распределения, построить гистограмму и полигон;
- получить точечные статистические оценки параметров распределения;
- построить теоретическую кривую и выдвинуть гипотезу о законе генерального распределения.

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									
-2,2	5,3	10,0	13,0	11,8	12,0	20,3	5,9	19,6	10,1
4,5	16,6	11,6	10,2	14,2	12,9	21,9	3,5	12,9	7,5
19,2	6,7	10,3	22,3	5,1	16,0	18,1	1,2	9,9	10,2
10,4	7,2	8,7	9,2	9,9	19,7	9,6	12,7	0,7	15,1
14,5	16,5	9,4	-0,2	4,8	2,8	15,2	0,9	9,5	9,6
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

Лабораторная работа № 2

СТАТИСТИЧЕСКАЯ ПРОВЕРКА ИСТИННОСТИ ВЫДВИНУТОЙ ГИПОТЕЗЫ. ИНТЕРВАЛЬНЫЕ ОЦЕНКИ ПАРАМЕТРОВ РАСПРЕДЕЛЕНИЯ

Цель работы: привить навыки по овладению методам проверки статистических гипотез, в частности, о нормальном законе распределения изучаемой случайной величины, признака, процесса; вычисление доверительных интервалов параметров нормального распределения, в которых с заданной вероятностью находятся соответствующие числовые характеристики генеральной совокупности.

Используемые программные средства: MS Excel 2010 (2016), Statistica 8.0.

2.1. Краткие теоретические сведения

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 $k_{\text{кр}}$.

После того как критическая точка найдена, по данным выборки вычисляют наблюдаемое значение критерии. Если $K_{\text{набл}} > k_{\text{кр}}$ (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 -значения, которое соответствует вероятности данного события при предположении, что некоторое утверждение (нулевая гипотеза) истинно. Если -значение меньше заданного уровня статистической значимости (обычно это 0,05) – нулевая гипотеза неверна, поэтому нужно перейти к рассмотрению альтернативной гипотезы.

Проверка гипотезы о распределении. Критерий Пирсона. Пусть $x_1, x_2, \dots, x_n$ – 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 $x_1, x_2, \dots, x_n$ 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 $a,\ \sigma$ и т.д.)

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

5. Рассчитывают *теоретические частоты* $n_i' = np_i$, где n – объем выборки.

6. Рассчитывают значение критерия согласия Пирсона

$$\chi^2 = \sum_{i=1}^{l} \frac{n_i - n_i'^{\,2}}{n_i'} \qquad (2.1)$$

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 $k = l - 3$. В дальнейшем для расчетов используют таблицы распределения χ^2.

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 альтер-

нативная 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 $np_i \geq 5$. Поэтому тот интервал, для которого это условие не выполняется, объединяют с соседним и соответственно уменьшают число интервалов.

Замечание 2.1. 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 $x \pm 0{,}3\sigma_B$ содержит примерно $\frac{1}{4}$ часть всей совокупности значений;
- промежуток $x \pm 0{,}7\sigma_B$ содержит примерно $\frac{1}{2}$ часть;
- промежуток $x \pm 1{,}1\sigma_B$ содержит примерно $\frac{3}{4}$ часть;
- промежуток $x \pm 3\sigma_B$ содержит примерно 0,99 всех значений.

Если эти соотношения выполняются одновременно для данной эмпирической совокупности и вычисленных x, σ_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 $\lambda = \overline{n} \cdot max_{x_i} F\ x_i - F^*\ x_i$
3) По уровню значимости α из таблицы 2.1 найти критическую точку $\lambda_{\text{кр}}$.
4) Если $\lambda < \lambda_{\text{кр}}$, то различия между эмпирическим и предполагаемым теоретическим распределениями несущественны. Если $\lambda > \lambda_{\text{кр}}$, то различия между эмпирическим и предполагаемым теоретическим распределениями существенны.

Таблица 2.1

α	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.001
$\lambda_{\text{кр}}$	1.138	1.2238	1.3581	1.4802	1.5174	1.6276	1.7308	1.9495

Значения точечных оценок могут значительно отличаться от соответствующих характеристик генеральной совокупности (особенно при малых объемах выборки). Для правильной оценки этого отклонения используют *ме-*

тод доверительных интервалов.

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 получается более широкий и менее полезный интервал. Если доверительная вероятность не задана, считают, что она равна 0,95 или 95%.

Среднее генеральной совокупности, имеющей нормальный закон распределения с доверительной вероятностью $\gamma = 1 - \alpha$, находится в доверительном интервале:

Точечной оценкой математического ожидания является выборочная средняя x. Границы доверительного интервала определяются как $a_1 = x - \delta$; $x + \delta$, где $\delta > 0$ – точность доверительного интервала, которая либо задается заранее, либо вычисляется.

Пусть известно, что случайная величина, из которой получена выборка объема n, имеет нормальный закон распределения. Истинное значение дисперсии этой случайной величины будем считать неизвестным. Рассмотрим два случая.

1) Если число наблюдений достаточно велико $n > 30$, то с вероятностью 0,95 значение математического ожидания попадает в доверительный интервал $x \pm 2s_x$, а с вероятностью 0,99 – в интервал $x \pm 3s_x$, где величина

$$s_x = \frac{s}{\sqrt{n}}$$

называется *стандартной ошибкой* (*ошибкой среднего*).

2) Если число наблюдений мало $n \leq 30$, то доверительный интервал определяется по формуле $x \pm \delta$, где

$$\delta = t\ \alpha, n-1\ \cdot \frac{s}{\sqrt{n}}$$

где $t\ \alpha, n-1$ – критическая точка распределения Стьюдента (для двусторонней критической области) с числом степеней свободы $n-1$ и уровнем значимости α.

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

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

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

Построим доверительный интервал для неизвестной дисперсии нормально распределенной генеральной совокупности. Оценкой для генеральной дисперсии является выборочная дисперсия. Доверительный интервал находится по следующей формуле:

$$\frac{n-1\ \cdot s^2}{\chi^2\ \ \frac{\alpha}{2}, n-1} < \sigma^2 < \frac{n-1\ \cdot s^2}{\chi^2\ \ 1-\frac{\alpha}{2}, n-1}$$

Значения $\chi^2\ \ \frac{\alpha}{2}, n-1$ и $\chi^2\ \ 1-\frac{\alpha}{2}, n-1$ находятся по таблицам хи-квадрат распределения.

Доверительный интервал для стандартного отклонения имеет вид:

$$\frac{\sqrt{n-1}}{\chi\ \frac{\alpha}{2}, n-1} \cdot s < \sigma < \frac{\sqrt{n-1}}{\chi\ 1-\frac{\alpha}{2}, n-1} \cdot s \qquad (2.3)$$

Или по одной из формул:

$$s \cdot\ 1-q\ \ < \sigma < s \cdot\ 1+q\ \text{, если } q < 1;$$
$$0 < \sigma < s \cdot\ 1+q\ \text{, если } \ q > 1.$$

Величина q определяется по таблице доверительных интервалов для σ по доверительной вероятности $\gamma = 1-\alpha$ и объёму выборки n.

2.2. Практическая часть.

Контрольный пример. Из генеральной совокупности извлечена случайная выборка (см. лабораторную работу № 1).

13,6	5,9	8,2	9,4	3,5	5,1	10,2	16,5	13,8	16,3
10,3	8,2	12,8	8,3	2,2	15,7	1,3	12,6	18,1	21,5
10,9	10,7	11,9	5,2	12,2	17,9	10	6,4	13	10,4
6	20,8	9,1	13,1	14,2	7,4	13,4	4,2	5,7	12,6
12,2	6,5	6,8	15,2	15,4	16,7	9,8	7,9	9,6	13,4

Требуется:

- провести приближенную проверку на нормальность;
- проверить, согласуются ли выборочные данные с гипотезой о нормальном распределении с помощью критериев Пирсона (в пакетах *Excel*, *STATISTICA*) и Колмогорова (в пакете *Excel*);

- найти доверительные интервалы для параметров нормального распределения (в пакете *STATISTICA*).

Принять $\alpha = 0{,}05$.

Решение. Поскольку (в предыдущей лабораторной работе) по виду гистограммы было выдвинуто предположение о нормальном распределении генеральной совокупности, то это предположение – основная выдвинутая гипотеза H_0.

Конкурирующая гипотеза H_1: генеральное распределение не является нормальным.

1. *Приближенная проверка с использованием* $\sigma_{\text{В}}$.

Вычислим (в пакете *Excel*) выборочное среднее x и выборочное СКО $\sigma_{\text{В}}$ с помощью стандартных функций СРЗНАЧ() и СТАНДОТКЛОН.Г():

	A	B	C	D
1	1,3		**xcp**	**10,842**
2	2,2		**σ**	**4,578**
3	3,5			

Вычислим значения:

$$0{,}3 \cdot \sigma_{\text{В}} = 0{,}3 \cdot 4{,}578 = 1{,}3734;\ \ 0{,}7 \cdot \sigma_{\text{В}} = 0{,}7 \cdot 4{,}578 = 3{,}2046;$$
$$1{,}1 \cdot \sigma_{\text{В}} = 1{,}1 \cdot 4{,}578 = 5{,}0358;\ \ 3 \cdot \sigma_{\text{В}} = 3 \cdot 4{,}578 = 13{,}374.$$

Вычислим границы интервалов:

$$x - 0{,}3 \cdot \sigma_B;\ x + 0{,}3 \cdot \sigma_B = 10{,}842 - 1{,}3734; 10{,}842 + 1{,}3734 = 9{,}5; 12{,}2\ ;$$
$$x - 0{,}7 \cdot \sigma_B;\ x + 0{,}7 \cdot \sigma_B = 10{,}842 - 3{,}2046; 10{,}842 + 3{,}2046 = 7{,}6; 14{,}1\ ;$$
$$x - 1{,}1 \cdot \sigma_B;\ x + 1{,}1 \cdot \sigma_B = 10{,}842 - 5{,}0358; 10{,}842 + 5{,}0358 = 5{,}8; 15{,}9\ ;$$
$$x - 3 \cdot \sigma_B;\ x + 3 \cdot \sigma_B = 10{,}842 - 13{,}374; 10{,}842 + 13{,}374 = -3; 24{,}6\ .$$

Подсчитаем число значений (из общей совокупности), попавших в вычисленные интервалы:

$$n_1 = 12;\ n_2 = 26;\ \ n_3 = 36;\ \ n_4 = 50$$

Вычислим относительные частоты:

$$\frac{n_1}{n} = \frac{12}{50} = 0{,}24;\ \ \frac{n_2}{n} = \frac{26}{50} = 0{,}52;\ \ \frac{n_3}{n} = \frac{36}{50} = 0{,}72;\ \frac{n_4}{n} = \frac{50}{50} = 1$$

Убеждаемся, что во втором и четвертом интервалах содержатся не менее рекомендуемого количества случайных чисел. А в первом и третьем – количе-

ство чисел близко к рекомендуемому.

Данное эмпирическое распределение скорее всего подчиняется нормальному закону распределения, но нужно провести проверку, используя более точные критерии.

2. *В пакете Statistica*.

После ввода исходных данных будем использовать процедуру *Distribution Fitting* (подбор распределения).

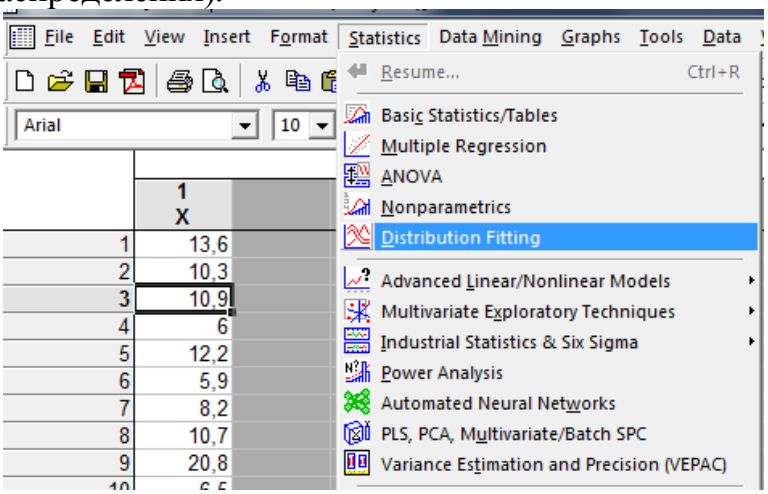


Рис. 2.1 – Вызов процедуры *Distribution Fitting*

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 нажмём кнопку *OK*.

Во вкладке *Parameters* установим количество интервалов разбиения равное 7. В этом же окне наблюдаем значения нижней и верхней границы значений исходных данных, наблюдаемые значение математического ожидания и дисперсии.

Нажав кнопку *Summary*, получаем таблицу частот (рис. 2.2).

Variable: X, Distribution: Normal (Spreadsheet1)
Chi-Square = 0,11262, df = 2 (adjusted) , p = 0,94525

Upper Boundary	**Observed Frequency**	Cumulative Observed	Percent Observed	Cumul. % Observed	Expected Frequency	Cumulative Expected	Percent Expected	Cumul. % Expected	Observed-Expected
<= 1,71429	1	1	2,00000	2,0000	1,21074	1,21074	2,42149	2,4215	-0,21074
5,42857	5	6	10,00000	12,0000	4,83444	6,04518	9,66887	12,0904	0,16556
9,14286	12	18	24,00000	36,0000	11,78811	17,83329	23,57621	35,6666	0,21189
12,85714	15	33	30,00000	66,0000	15,59054	33,42383	31,18109	66,8477	-0,59054
16,57143	12	45	24,00000	90,0000	11,19074	44,61457	22,38148	89,2291	0,80926
20,28571	3	48	6,00000	96,0000	4,35640	48,97097	8,71280	97,9419	-1,35640
< Infinity	2	50	4,00000	100,0000	1,02903	50,00000	2,05806	100,0000	0,97097

Рис. 2.2 – Таблица частот

Если гипотеза верна, вероятность получить 0,11262 или больше равна 0.94525 (больше 0,05 – уровня значимости) – достаточна, чтобы поверить в

нормальность распределения исходных данных. Следовательно, гипотезу о нормальном распределении случайной величины принимаем.

Кроме того, нормальность распределения приближенно можно оценить графически по нормальным вероятностным графикам (*Graphs-2D Graphs-Normal Probability Plots*): чем ближе опытные точки к прямой линии, тем ближе распределение к нормальному (рис. 2.3).

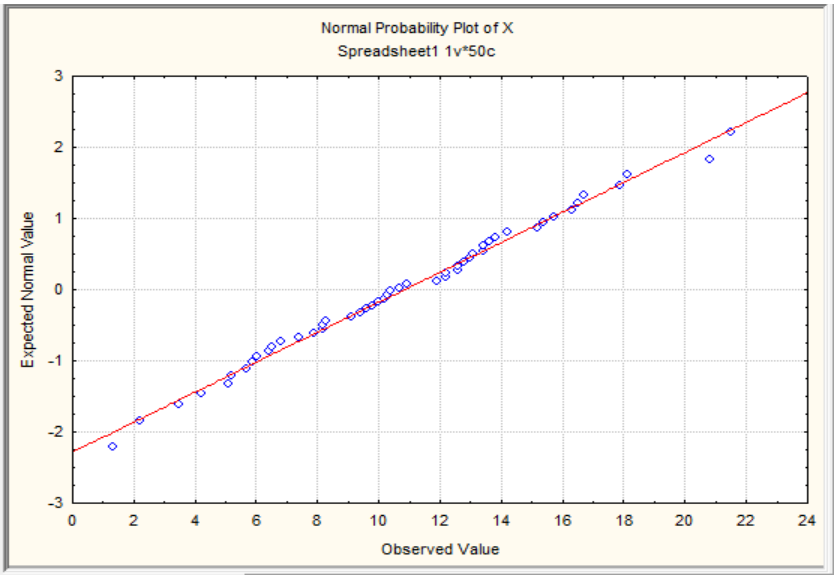


Рис. 2.3 – Нормальный вероятностный график

3. В пакете *Excel*:

Введём исходные данные и оформим его так, как показано на рис. 2.4. Вычислим выборочные характеристики, используя стандартные функции пакета.

	A	B	C	D	E	F	G	H
1	Наблюдения		Числовые характеристики			Левые границы интервалов	Правые границы интервалов	Эмпирические частоты
2	1,3		среднее	10,8420		1,3	4,186	3
3	2,2		станд.откл.	4,6250		4,186	7,072	9
4	3,5		асимметрия	0,1340		7,072	9,958	9
5	4,2		эксцесс	-0,3232		9,958	12,844	12
6	5,1					12,844	15,73	10
7	5,2					15,73	18,616	5
8	5,7					18,616	21,502	2
9	5,9					Всего наблюдений		50

Рис. 2.4 – Расчет основных выборочных характеристик выборки

В ячейку I2 внесём формулу для вычисления значения функции нормального распределения $F\ x_1 = 4{,}186$. В *Excel* эту величину можно вычислить, воспользовавшись функцией НОРМ.РАСП (рис. 2.5).

В поле X введён адрес ячейки, в которой находится граница первого интервала группировки.

В поле *Среднее* введён адрес ячейки, в которой находится среднее значение выборки.

В поле *Стандартное_откл* введён 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).

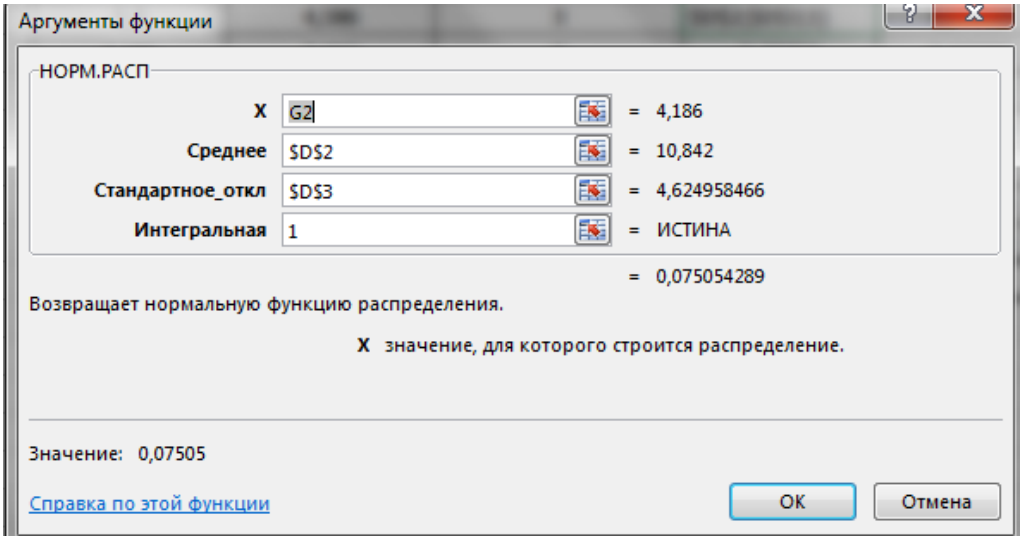


Рис. 2.5 – Диалоговое окно функции НОРМРАСП с заполненными полями ввода

Далее с помощью маркера автозаполнения протянем эту формулу до ячейки I8.

В ячейку J2 введём формулу =I2, в ячейку J8 формулу =1-I7, а в ячейку J3 формулу =I3-I2, протянув её до ячейки J7 (рис. 2.6). В результате этих действий в диапазоне J2: J8 появятся значения теоретических вероятностей $p_1, p_2, \dots, p_7$, причём $p_1 + p_2 + \cdots + p_7 = 1$.

B	C	D	E	F	G	H	I	J
	Числовые характеристики			Левые границы интервалов	Правые границы интервалов	Эмпирические частоты	Функция нормального распределения	Теоретические вероятности
	среднее	10,8420		1,3	4,186	3	0,07505	0,07505
	станд.откл.	4,6250		4,186	7,072	9	0,20750	0,13244
	асимметрия	0,1340		7,072	9,958	9	0,42421	0,21671
	эксцесс	-0,3232		9,958	12,844	12	0,66744	0,24324
				12,844	15,73	10	0,85472	0,18727
				15,73	18,616	5	0,95361	0,09889
				18,616	21,502	2	0,98941	0,04639
				Всего наблюдений		50		1

Рис. 2.6 – Расчёт теоретических вероятностей

В ячейку K2 введём формулу = $H$9*J2 и протянем её до ячейки K8. С

помощью этой формулы вычисляются теоретические частоты n_i'. Их сумма равна объему выборки $n = 50$. На этом заканчивается первый этап проверки (рис. 2.7).

F	G	H	I	J	K
Левые границы интервалов	Правые границы интервалов	Эмпирические частоты	Функция нормального распределения	Теоретические вероятности	Теоретические частоты
1,3	4,186	3	0,07505	0,07505	3,75271
4,186	7,072	9	0,20750	0,13244	6,62205
7,072	9,958	9	0,42421	0,21671	10,83569
9,958	12,844	12	0,66744	0,24324	12,16179
12,844	15,73	10	0,85472	0,18727	9,36353
15,73	18,616	5	0,95361	0,09889	4,94457
18,616	21,502	2	0,98941	0,04639	2,31966
Всего наблюдений		50		1	50

Рис. 2.7 – Первый этап проверки гипотезы по критерию хи-квадрат

По данным в ячейках H2:H8 и K2: K8 построим диаграмму фактических (эмпирических) и теоретических частот. Сначала скопируем на новый лист пакета данные, расположенные в столбцах H и K (рис. 2.8).

	A	B	C
1	Варианты	Эмпирические частоты	Теоретические частоты
2	2,743	3	3,75271
3	5,629	9	6,62205
4	8,515	9	10,83569
5	11,401	12	12,16179
6	14,287	10	9,36353
7	17,173	5	4,94457
8	20,059	2	2,31966

Рис. 2.8 – Исходные данные для построения диаграммы

Здесь в столбце *Варианты* находятся середины интервалов (см. лабораторную работу № 1).

Чтобы совместить в диаграмме несколько типов, (например, Гистограмму и Линейный график – как в нашем примере), необходимо сначала построить все диаграммы одного вида. Выделим диапазон B2: C8 и на вкладке «Вставка» выберем тип Гистограмма (рис. 2.9 – 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 выбираем тип «График» (рис. 2.9 – б).

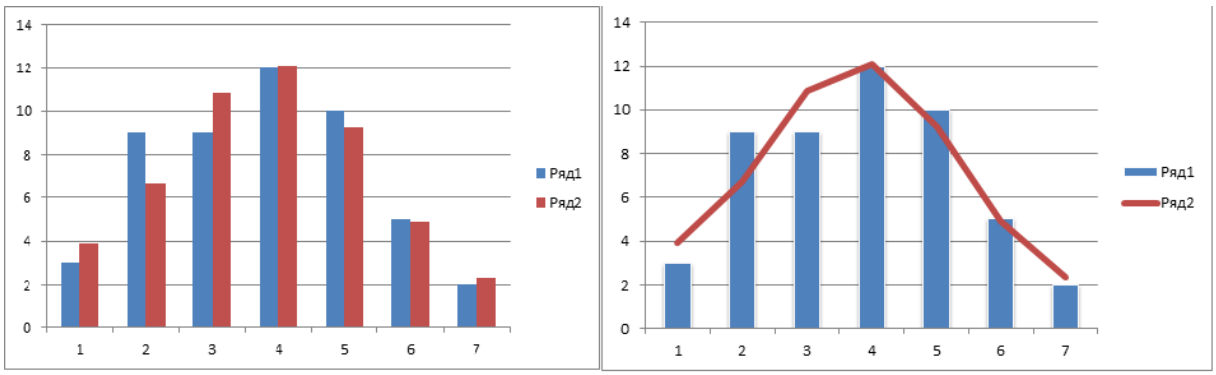


Рис. 2.9 – Первый этап построения диаграммы

Добавим вспомогательную ось. Для этого нажмём 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. 2.10.

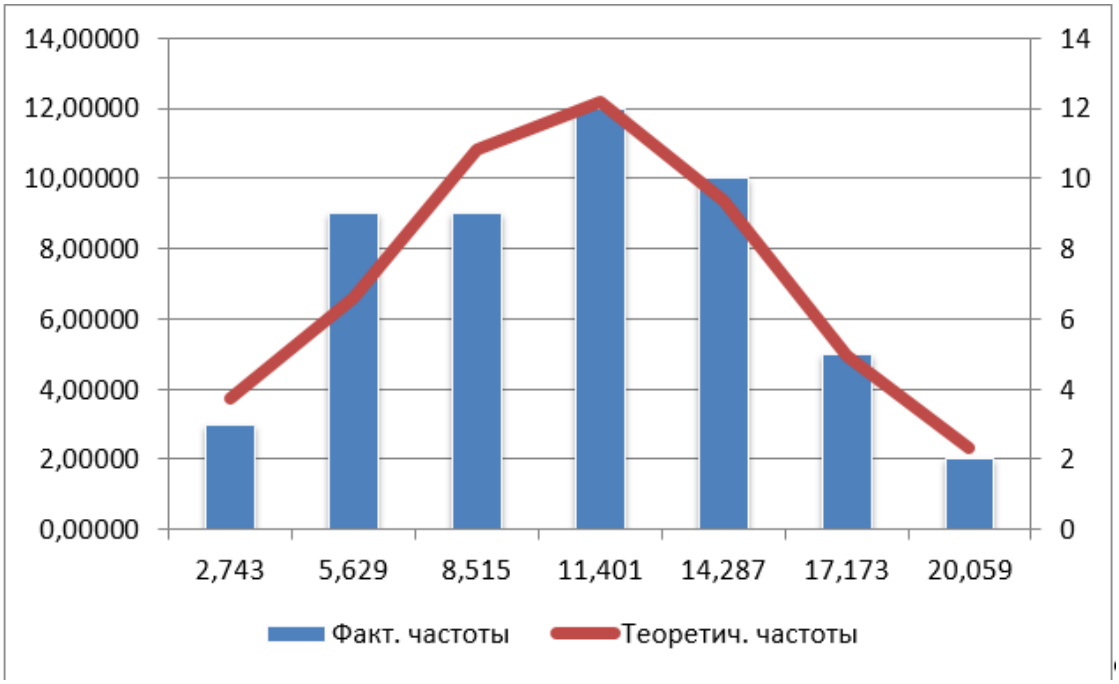


Рис. 2.10 – Диаграмма фактических и теоретических частот

Визуальное сравнение гистограммы и графика теоретических частот дает основание полагать, что исследуемое распределение близко нормальному. График исследуемого распределения является симметричным (асимметрия мала), но более пологим, чем график теоретического нормального распределения (эксцесс имеет небольшое отрицательное значение) (см. рис. 2.4).

Результаты заключительного этапа проверки приведены в диапазоне M1:O10 (рис. 2.12).

В диапазоне M2:M7 находятся групповые частоты n_i, в диапазоне N2:N7 – ожидаемые частоты n_i'.

Так как ожидаемые частоты первого и седьмого интервалов группировки не удовлетворяют условию $\min_i\ n_i \geq 5$, эти интервалы объединены со вторым и шестым интервалами.

После такого объединения число l интервалов группировки становится равно 5.

В ячейку O2 введена формула =(M2-N2)^2/N2, реализующая вычисления по формуле $\frac{n_i - n_i'^{\,2}}{n_i'}$. Размножим эту формулу в диапазоне ячеек O3:O6. В ячейке N9 получим сумму содержимого ячеек O2:O6.

Критическое значение статистики U, которая имеет распределение с двумя (число частичных интервалов – 5; 5 – 3 = 2) степенями свободы, определяется при помощи функции ХИ2.ОБР.ПХ (рис. 2.11).

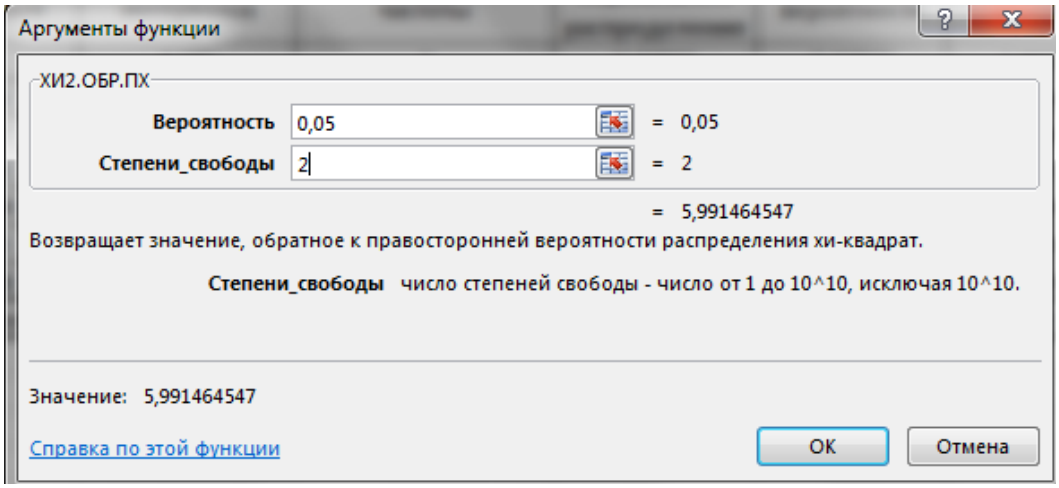


Рис. 2.11 – Диалоговое окно функции ХИ2ОБР с заполненными полями ввода

Расчётное значение $\chi^2_{\text{набл}} = 0{,}6206$ статистики U меньше её критического значения $\chi^2_{\text{кр}} = 5{,}9915$ (рис. 2.12), поэтому можно сказать, что проверяемая гипотеза, состоящая в том, что генеральная совокупность подчиняется нормальному закону распределения, не противоречит данным эксперимента.

F	G	H	I	J	K	M	N	O
Левые границы интервалов	Правые границы интервалов	Эмпирические частоты	Функция нормального распределения	Теоретические вероятности	Теоретические частоты	Эмп. Частоты	Теор. Частоты	U
1,3	4,186	3	0,07505	0,07505	3,75271	12	10,37477	0,254596003
4,186	7,072	9	0,20750	0,13244	6,62205	9	10,83569	0,310986944
7,072	9,958	9	0,42421	0,21671	10,83569	12	12,16179	0,00215225
9,958	12,844	12	0,66744	0,24324	12,16179	10	9,36353	0,043263003
12,844	15,73	10	0,85472	0,18727	9,36353	7	7,26422	0,009610659
15,73	18,616	5	0,95361	0,09889	4,94457	50	50	
18,616	21,502	2	0,98941	0,04639	2,31966	Наблюдаемое значение Хи-квадрат		
Всего наблюдений		50		1	50		0,6206	
						Критическое значение Хи-квадрат		
							5,9915	

Рис. 2.12 – Таблица с окончательными результатами вычисления статистики

4. *Проверка истинности гипотезы* H_0 *по критерию Колмогорова*.

Скопируем данные наблюдений на чистый лист *Excel* и построим интервальный вариационный ряд, аналогично тому, как делалось в лабораторной работе № 1 (см. рис. 2.13).

Рассчитаем выборочную среднюю и «исправленное» СКО, используя функции СРЗНАЧ(A2:A51) и СТАНДОТКЛОН.В(A2:A51) – результат в ячейках D4 и D5.

	A	B	C	D	E	F	G	H	I	J	K	L
1	**Наблюдения**		**Числовые характеристики**		**xi-1**	**xi**	**Частоты**	**Относительные частоты**	**Накопленные отн. частоты (F*(x))**	**(xi-xs)/s**	**F(xi)**	**|F*(x)-F(x)|**
2	13,6		**максимум**	**21,5**	менее 1,3		0	0	0	-2,3442	0,0095	0,0095
3	10,3		**минимум**	**1,3**	1,3	4,186	3	0,06	0,06	-1,4391	0,0751	0,0151
4	10,9		**среднее xs**	**10,842**	4,186	7,072	9	0,18	0,24	-0,8151	0,2075	0,0325
5	6		**станд_откл s**	**4,625**	7,072	9,958	9	0,18	0,42	-0,1911	0,4242	0,0042
6	12,2				9,958	12,844	12	0,24	0,66	0,4329	0,6674	0,0074
7	5,9				12,844	15,73	10	0,2	0,86	1,0569	0,8547	0,0053
8	8,2				15,73	18,616	5	0,1	0,96	1,6809	0,9536	0,0064
9	10,7				18,616	21,502	2	0,04	1	2,3049	0,9894	0,0106
10	20,8			**Всего наблюдений**			**50**	**1**				
11	6,5											
12	8,2								**Наибольшее отклонение**			**0,0325**
13	12,8								**Наблюдаемое значение статистики**			**0,2298**
14	11,9								**Критическая точка для α=0,05**			**1,3581**
15	9,1											
16	6,8								***Нет оснований отклонить гипотезу H0***			

Рис. 2.13 – Расчёты для критерия Колмогорова

В столбцах H и I рассчитаем относительные и накопленные частоты (см. лабораторную работу № 1). Накопленные частоты (точки кумулятивной кривой) – левые концы «ступенек» эмпирической функции распределения, т.е. $F^* \ x$.

Значения $F \ x_i$ (см. столбец K) вычисляются с учётом того, что была выдвинута гипотеза о нормальном распределении, т.е.

$$F \ x_i \ = 0{,}5 + \Phi \ \frac{x_i - x}{s}$$

В пакете Excel значения $F \ x_i$ можно вычислить с помощью функции НОРМ.СТ.РАСП():

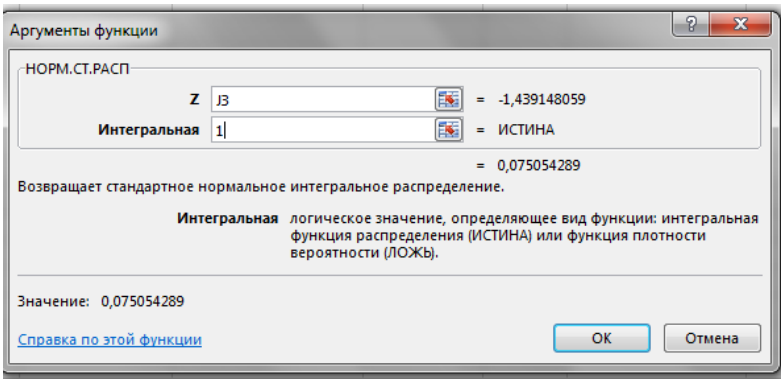


Рис. 2.14 – Диалоговое окно функции НОРМ.СТ.РАСП() с заполненными полями ввода

В столбце L рассчитаем абсолютное значение отклонения эмпирической

функции распределения от теоретической, при этом из последнего столбца таблицы ясно, что $max_i\ F^*\ x_i\ - F\ x_i\ = 0{,}0325$.

В ячейке L13 найдём значение статистики Колмогорова

$$\lambda = \sqrt{n} \cdot max_{x_i}\ F\ x_i\ - F^*\ x_i\ = \sqrt{50} \cdot 0{,}0325 \approx 0{,}2298$$

По заданному уровню значимости $\alpha = 0{,}05$ определим границу критической области (табл. 2.1) $\lambda_{kp} = 1{,}3581$. Поскольку $\lambda = 0{,}2298 < \lambda_{kp} = 1{,}3581$, то гипотезу H_0 о том, что закон распределения генеральной совокупности является нормальным, отвергнуть нельзя.

5. *Интервальные оценки параметров распределения*

Введем исходные данные (рис. 2.15). Для того чтобы найти доверительные интервалы для математического ожидания при $\gamma = 0{,}95$ добавим два столбца и назовём их xs и s.

	1 x	2 xs	3 s
1	1,3	10,842	4,625
2	2,2		
3	3,5		

Рис. 2.15 – Добавление столбцов xs и s

Значения выборочной средней и среднего квадратического отклонения были подсчитаны ранее (в пакете *Excel*).

В меню *Statistics* активизируем модуль *Basic Statistics and Tables*, в котором выберем процедуру *Probability Calculator*. В поле Distribution слева выберем распределение Стьюдента t(Student), $p = 1 - \frac{\alpha}{2} = 1 - \frac{0{,}05}{2} = 0{,}975$; $df = n - 1 = 49$. Далее нужно щелкнуть левой кнопкой на *Compute* и в поле t появится критическая точка распределения Стьюдента $t_{\text{кр}}\ \alpha; n-1$ для двусторонней критической области (рис. 2.16).

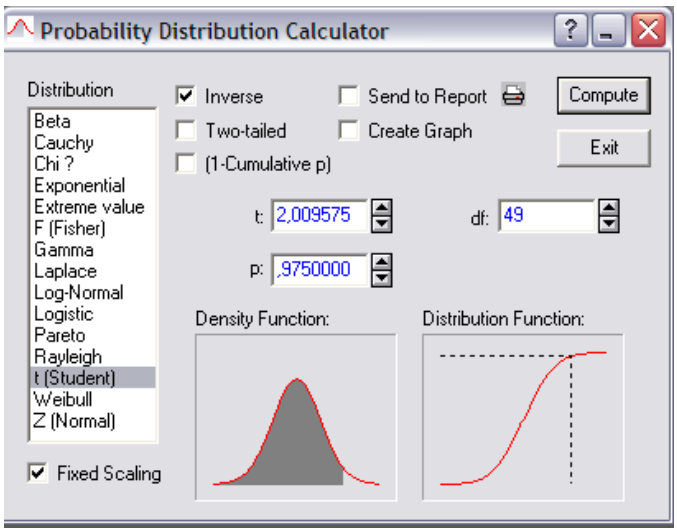


Рис. 2.16 – Окно вероятностного калькулятора

Нажмем кнопку *Compute*; результат в поле t (рис. 2.16). Выйдем из модуля, нажав кнопку *Exit*.

Определим столбцы a_1 и a_2 левых и правых концов доверительного интервала.

Добавим два столбца для левой и правой границы доверительного интеграла для среднего. Выделим заголовок столбца xs. Нажмем на правую кнопку мыши, выберем *Add Variables*, установим *How many*: 2; *after*: s. Нажмем OK. Выделим новый столбец, нажмем правую кнопку мыши, выберем команду *Variable Specs*, установим следующие параметры: имя *Name: A1* (левые концы), *Long name:* $= xs - 1{,}64854 * s \,/\, Sqrt(50)$ (рис. 2.17, рисунок слева).

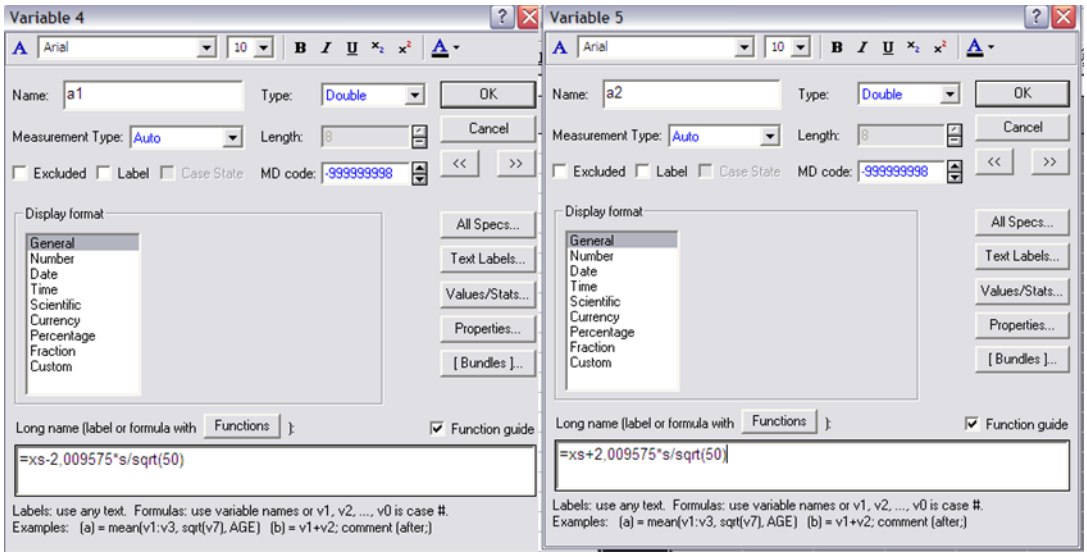


Рис. 2.17 – Задание параметров для вычисления границ доверительных интервалов

После *OK* получаем левую границу доверительного интервала. Аналогично получаем правую границу a_2 (рис. 2.17, рисунок справа).

Вычисленные результаты приведены на рис. 2.18

	1 x	2 xs	3 s	4 a1	5 a2
1	1,3	10,842	4,625	9,52759	12,15641
2	2,2				

Рис. 2.18 – Вычисление доверительного интервала для математического ожидания

Для построения доверительного интервала для дисперсии в исходную таблицу добавим 4 столбца, назовем их *n*, *D*, *D1*, *D2*. (рис. 2.19)

	1 x	2 xs	3 s	4 a1	5 a2	6 n	7 D	8 D1	9 D2
1	1,3	10,842	4,625	9,52759	12,15641	50	21,39063		
2	2,2								

Рис. 2.19 – Добавление столбцов количества измерений и дисперсии

Значения левой и правой границ доверительного интервала будем вычислять по формуле $\frac{n-1}{\chi_2^2} < D < \frac{n-1}{\chi_1^2}$. Величины χ_1^2 и χ_2^2 – это значения распределения χ^2 в точках $1+\gamma\ /2$ и $1-\gamma\ /2$ с количеством степеней свободы $n-1$, γ – доверительная вероятность. В данной задаче $n = 50$, $\gamma = 0{,}95$. Зададим расчетные формулы для переменных D1 и D2:

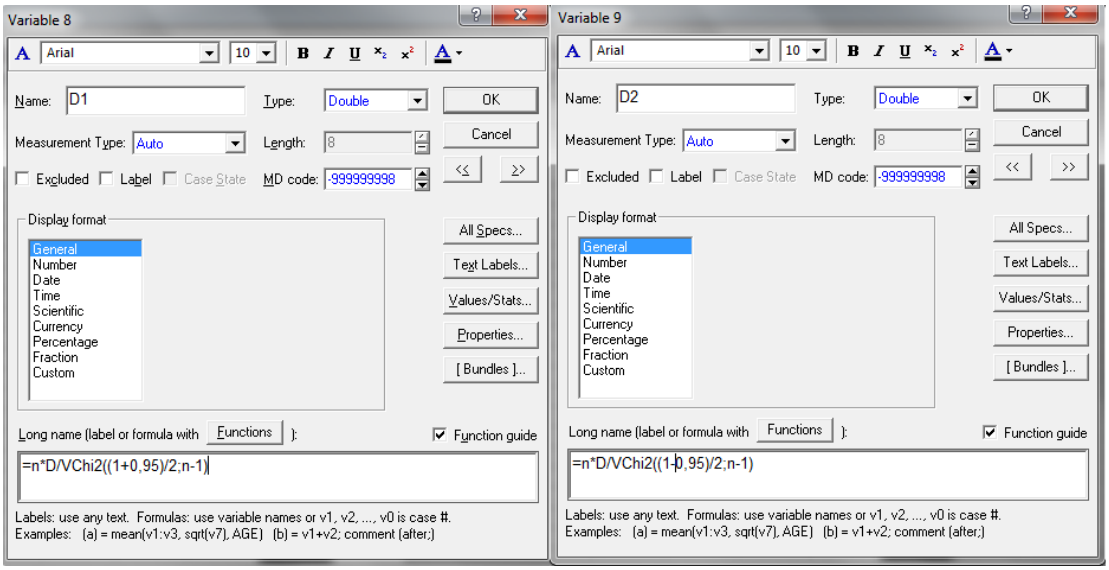


Рис. 2.20 – Расчет границ доверительного интервала

На рис. 2.20 слева задана формула для левой границы доверительного интервала, справа – для правой. Пересчитав переменные, получим требуемый результат.

	1 x	2 xs	3 s	4 a1	5 a2	6 n	7 D	8 D1	9 D2
1	1,3	10,842	4,625	9,52759	12,15641	50	21,39063	15,23063	33,89428
2	2,2								

Рис. 2.21 – Результаты вычислений

Доверительный интервал для среднего квадратичного отклонения имеет вид:

$$3{,}9026 < \sigma < 5{,}8219.$$

2.3. Задание для самостоятельной работы.

Используя данные своего варианта из лабораторной работы 1:

- провести приближенную проверку на нормальность;
- проверить, согласуются ли выборочные данные с гипотезой о нормальном распределении с помощью критериев Пирсона (в пакетах *Excel*, *STATISTICA*) и Колмогорова (в пакете *Excel*);
- найти доверительные интервалы для параметров нормального распределения (в пакете STATISTICA).

Принять $\alpha = 0{,}05$.

Лабораторная работа № 3

ПРОВЕРКА ГИПОТЕЗ О ПАРАМЕТРАХ РАСПРЕДЕЛЕНИЙ

Цель работы: привить навыки по овладению методам проверки статистических гипотез о параметрах распределения нормально распределенной случайной величины.

Используемые программные средства: MS Excel 2010 (2016), STATISTICA 8.0.

3.1. Краткие теоретические сведения.

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

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 $\mathrm{H}_0: a = a_0;\ \mathrm{H}_1: a \neq a_0\ \ a < a_0; a > a_0$.

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

$$T = \frac{x - a_0}{s} \cdot \ \overline{n} \qquad (3.1)$$

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

Для того чтобы при заданном уровне значимости α проверить $\mathrm{H}_0: a = a_0$ при альтернативной гипотезе $\mathrm{H}_1: a > a_0$, по таблице распределения Стьюдента находят квантили $t_{kp} = t\ \alpha, k$ из равенства $P\ \ T > t\ \alpha, k\ \ = \alpha$.

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

При альтернативной гипотезе $\mathrm{H}_1: a < a_0$, по таблице распределения Стьюдента находят квантиль $t_{kp} = t\ \alpha, k$ из равенства $P\ \ T > t\ \alpha, k\ \ = 1 - \alpha$.

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

При альтернативной гипотезе $\mathrm{H}_1: a \neq a_0$, сравнивают модуль статистической характеристики T с квантилем $t_{kp} = t\ \alpha, k$ распределения Стьюдента, найденным из равенства $P\ \ T \leq t\ \ \frac{\alpha}{2}, k\ \ = P\ \ T \geq t\ \ \frac{\alpha}{2}, k\ \ = \frac{\alpha}{2}$.

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

Замечание 3.1. В пакете MS Excel квантиль распределения Стьюдента модно нацти с помощью стандартных функций СТЬЮДЕНТ.ОБР.2Х($\alpha;k$) для двусторонней критической области и СТЬЮДЕНТ.ОБР($1-\alpha;k$) – для односторонней; в пакете STATISTICA – с помощью вероятностного калькулятора.

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 генеральной 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

$$\chi^2 = \frac{n-1\ s^2}{\sigma_0^2} \tag{3.2}$$

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 $P\ \ \chi^2 > \chi^2_{kp}\ \ \alpha, k\ \ = \alpha$, а критическую точку ищут по таблицам квантилей распределения χ^2. После этого вычисляем по данной выборке наблюдаемое значение критерия. Если $\chi^2_{\text{набл}} < \chi^2_{kp}\ \ \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_{kp}\ \ 1-\alpha, k\ $. Тогда, если $\chi^2_{\text{набл}} > \chi^2_{kp}\ \ \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:

$$P\ \ \chi^2 < \chi^2_{\text{лев}}\ \ 1-\alpha/2, k\ \ \ = \alpha/2; P\ \ \chi^2 > \chi^2_{\text{прав}}\ \ \alpha/2, k\ \ \ = \alpha/2.$$

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

Задача сравнения дисперсий возникает при сравнении точности прибо-

ров, 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 $a_1,\ \sigma_1$ и $a_2,\ \sigma_2$. Из генеральных совокупностей выполнены выборки объёмами n_1 и n_2, и вычислены точечные оценки x, 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

$$F_{\text{набл}} = \frac{s_{\text{Б}}^2}{s_{\text{М}}^2} \tag{3.3}$$

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 (3.3). Критическое – находится по таблице критических точек распределения Фишера $F_{\text{кр}}\ \frac{\alpha}{2}, k_1, k_2\ ;$ (k_1 – число степеней свободы большей дисперсии).

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

2) $\text{H}_0: \sigma_1^2 = \sigma_2^2; \text{H}_1: \sigma_1^2 > \sigma_2^2$.

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:

$$P\ F > F_{\text{кр}}\ \alpha, k_1, k_2\ = \alpha$$

Наблюдаемое зачение критерия вычисляется по формуле (3.3). Критическое – находится по таблице критических точек распределения Фишера $F_{\text{кр}}\ \alpha, k_1, k_2\ .$

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

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 x и y, а также исправленные выборочные дисперсии s_x^2 и s_y^2.

1) Проверяемая гипотеза $\mathrm{H}_0: a_1 = a_2$, 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:

$$T = \frac{x-y}{\sqrt{n_1-1 \cdot s_x^2 + n_2-1 \cdot s_y^2}} \cdot \sqrt{\frac{n_1 \cdot n_2 \cdot n_1+n_2-2}{n_1+n_2}} \qquad (3.4)$$

В том случае, когда проверяемая гипотеза верна, статистика T, определяемая формулой (3.4), 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 (3.4);
- $k = n_1 + n_2 - 2$ – число степеней свободы;
- $t_{\text{двуст.кр}}\ \alpha, k$ – критическая точка двусторонней критической области – находится по таблице критических точек распределения Стьюдента.

При конкурирующей гипотезе $H_1: a_1 > a_2$ находят критическую точку $t_{\text{правост.кр}}\ \alpha, k$ по таблице критических точек распределения Стьюдента 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 $t_{\text{левост.кр}}\ \alpha, k = -t_{\text{правост.кр}}\ \alpha, k$. 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

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

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

$$k = \frac{\left(\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}\right)^2}{\frac{\left(\frac{s_x^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_y^2}{n_2}\right)^2}{n_2 - 1}} \tag{3.6}$$

Область принятия гипотезы для двусторонней критической области (альтернатива $\mathrm{H}_1: a_1 \neq a_2$) имеет вид:

$$\frac{x - y}{\sqrt{\frac{s_x^2}{n_1} + \frac{s_y^2}{n_2}}} < t_{\text{двуст.кр.}}\left(\frac{\alpha}{2}, k\right) \tag{3.7}$$

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,

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

$s_d = \sqrt{\frac{\sum_i d_i^2 - \frac{\left(\sum_i d_i\right)^2}{n}}{n-1}}$ – «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:

- вычислить наблюдаемое значение критерия $T_{\text{набл}} = \frac{d}{s_d}\sqrt{n}$
- по таблице критических точек распределения Стьюдента, по заданному уровню значимости α для двусторонней критической области и числу степеней свободы $k = n - 1$ найти критическую точку $t_{\text{двуст.крит.}}(\alpha, k)$;
- если $\mathrm{T}_{\text{набл}} < t_{\text{двуст.кр}}$ – нет оснований отвергать нулевую гипотезу. Если $\mathrm{T}_{\text{набл}} > t_{\text{двуст.кр}}$ – нулевую гипотезу отвергают.

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 $s_1^2, s_2^2, \dots, s_l^2$, 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:

$$D\ X_1\ = D\ X_2\ = \cdots = D\ X_l\ .$$

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:

$$G = \frac{s_{max}^2}{s_1^2+s_2^2+\cdots+s_l^2} \qquad (3.8)$$

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 $P\ G > G_{kp}(\alpha, k, l)\ = \alpha$.

Критическую точку $G_{kp}(\alpha, k, l)$ находят по таблице приложения (или с помощью стандартной функции пакета Excel БЕТА.ОБР$(1-\frac{\alpha}{l}; \frac{n-1}{2}; \frac{l\cdot(n-1)}{2})$, и тогда правосторонняя критическая область определяется неравенством $G > G_{kp}$, а область принятия нулевой гипотезы – $G < G_{kp}$.

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

3.2. Практическая часть

Контрольный пример 3.1. Проектный, контролируемый размер деталей, изготовляемых станком-автоматом, $a = a_0 = 35$ мм. Измерения n случайным образом отобранных деталей, дали следующие результаты:

Длина деталей, x_i	34.8	34.9	35	35.1	35.3
Частота (число деталей) n_i	2	3	4	6	5

Требуется:

а) при уровне значимости 0,05 проверить нулевую гипотезу $\mathrm{H}_0: a = 35$ (станок обеспечивает проектный размер деталей) при конкурирующей гипотезе $\mathrm{H}_1: a \neq 35$;

б) партия деталей принимается, если дисперсия контролируемого размера значимо не превышает $\sigma_0^2 = 0{,}2$. Можно ли принять партию при уровне значимости 1) 0,01; 2) 0,05? Принять в качестве альтернативной $\mathrm{H}_1: \sigma^2 > \sigma_0^2$.

*Решение.*1) Проверим гипотезу в пакете *Excel.*

В диапазон A2:A21 листа *Excel* введём исходные данные наблюдения (рис. 3.1).

D10 | =ЕСЛИ(ABS(D4)<D6;"станок обеспечивает проектный размер деталей";"нулевая гипотеза отвергается")

	A	B	C	D	E	F	G	H	I	J	K
1	Длина деталей		Проверка гипотезы о математическом ожидании								
2	34,8		Выб. Среднее	35,07							
3	34,8		s	0,1658							
4	34,9		T	1,8887							
5	34,9		Критич. точка для двусторонней критической области								
6	34,9		t(α; n-1)	2,0930							
7	35		Нахождение критической точки с помощью функции								
8	35		СТЬЮДЕНТ.ОБР(1-α/2;n-1)	2,0930							
9	35										
10	35		Вывод	станок обеспечивает проектный размер деталей							
11	35,1										

Рис. 3.1 – Проверка нулевой гипотезы о проектном размере детали

В ячейку D2 введём формулу $= \text{СРЗНАЧ}\ \text{A2:A21}$ и нажмём клавишу *Enter*. В ячейке появится средний размер деталей выборки

В ячейку D3 введём формулу $= \text{СТАНДОТКЛ.В(A1:A12)}$ и нажмём клавишу *Enter*. В ячейке появится «исправленное» выборочное стандартное отклонение.

В ячейку D4 введём формулу $= (\text{D2} - 35) * \text{КОРЕНЬ}(20)/\text{D3}$ и нажмём клавишу *Enter*.В ячейке появится расчётное значение статистики T.

Так как конкурирующая гипотеза имеет вид $\mathrm{H}_1: a \neq 35$, то критическая область – двусторонняя.

В ячейку D6 введём формулу $= \text{СТЬЮДЕНТ.ОБР.2Х}(0{,}05; 19)$ и нажмём клавишу *Enter*. В ячейке появится критическое значение распределения Стьюдента для двусторонней критической области с 19 степенями свободы.

Полученный результат $T < t_{kp}$ свидетельствует о том, что гипотеза $\mathrm{H}_0: a = 35$ не противоречит данным эксперимента – станок обеспечивает проектный размер деталей.

Замечание 3.2. Критическое значение можно найти с помощью стандартной функции =СТЬЮДЕНТ.ОБР(1-0,025;11) – результат будет аналогичным.

2) Проверим гипотезу в пакете *Statistica*.

Введём исходные данные:

	1 X		
1	34,8	11	35,1
2	34,8	12	35,1
3	34,9	13	35,1
4	34,9	14	35,1
5	34,9	15	35,1
6	35	16	35,3
7	35	17	35,3
8	35	18	35,3
9	35	19	35,3
10	35,1	20	35,3

Рис. 3.2 – Исходные данные

Для проверки гипотезы будем использовать процедуру *Basic Statistics/Tables*, которая находится в меню *Statistics*.

Выбираем *t-test, single sample*:

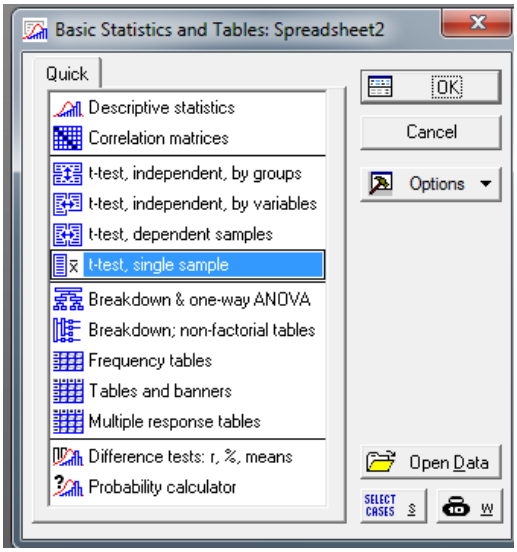


Рис. 3.3 – Процедура *Basic Statistics/Tables*

Зададим диапазон исходных данных, нажав на кнопку *Variables* и выбрав там X. Нажимаем кнопку *OK*. Далее на вкладке *Advanced* зададим исходные данные задачи (рис. 3.4):

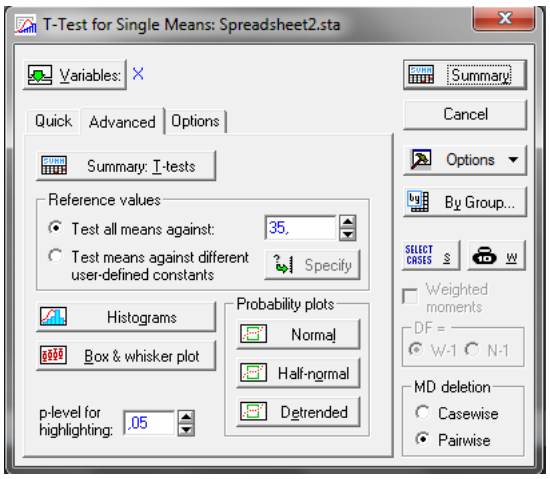


Рис. 3.4 – Исходные данные задачи на вкладке *Advanced*

Нажав кнопку *Summary*, получаем таблицу (рис.3.5):

Variable	Test of means against reference constant (value) (Spreadsheet2.sta)							
	Mean	Std.Dv.	N	Std.Err.	Reference Constant	t-value	df	p
X	35,07000	0,165752	20	0,037063	35,00000	1,888663	19	0,074301

Рис. 3.5 – Проверка гипотезы о математическом ожидании при неизвестной дисперсии

Так как статистическая значимость $p = 0{,}074301 > 0{,}05$, то приходим к выводу, что проверяемая гипотеза не противоречит опытным данным.

2) Проверяемая гипотеза эквивалентна гипотезе $\mathrm{H}_0: \sigma^2 = 0{,}2$. Примем в качестве конкурирующей гипотезы (согласно условию) $\mathrm{H}_1: \sigma^2 > 0{,}2$.

На рисунке 3.6 представлено решение данной задачи в пакете *Excel*.

H	I	J	K	L	M	N	O
	Проверка гипотезы о значении дисперсии						
	Оценка дисперсии s^2			0,0275			
	Наблюдаемое значение критерия			2,61			
	Квантиль распределения Хи-квадрат						
	для правосторонней крит. области			30,1435			
	Вывод						
	партия деталей принимается						

Рис. 3.6 – Проверка гипотезы о значении дисперсии в пакете *Excel*

В ячейке L2 с помощью стандартной функции пакета ДИСП. В(A2: A21) была найдена «исправленная» дисперсия выборки.

В ячейку L4 введена формула $= \mathrm{L2} * 19/0{,}3$, в ячейку L7 – функция ХИ2. ОБР. ПХ(0,05; 19).

Сравнивая расчётное значение статистики χ^2 с её критическим значением порядка 0.05, приходим к выводу, что проверяемая гипотеза не противоречит данным наблюдения – партия деталей принимается.

В пакете STATISTICA:

Введем исходные данные (рис. 3.7).

Добавим три столбца, назовем их $dispers, X\_H$ и X_kp. В столбец $dispers$ введем значение «исправленной» дисперсии, вычисленной ранее в пакете *Excel* (см. рис. 3.7).

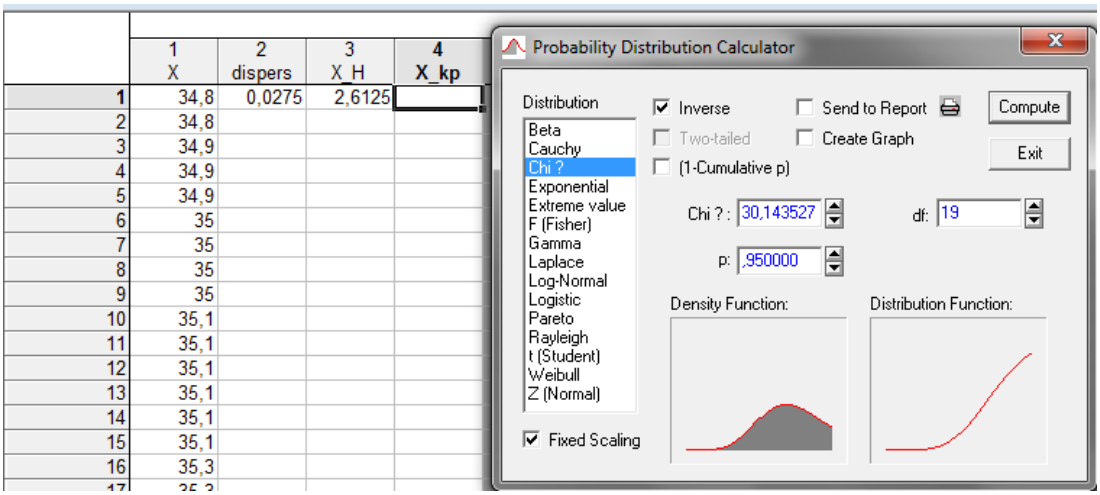


Рис. 3.7 – Проверка гипотезы о значении дисперсии в пакете *STATISTICA*

Щелкнем правой кнопкой мыши по названию столбца X_H и в поле $Long\ name$ введем формулу $= 19 * dispers/0{,}2$, где 19 – число степеней свободы, равное $n - 1 = 20 - 1 = 19$.

Критическое значение найдем с помощью вероятностного калькулятора (*Statistics – Probability Calculator – Distributions*). В поле $Distribution$ (см. рис. 3.8) выберем Chi ?, в поле p – 0,95 и нажмем кнопку $Compute$. Результат будет отображен в поле Chi ? (см. рис. 3.7).

Так как $\chi^2_{\text{набл}} = 2{,}1625 < \chi^2_{kp} = 30{,}1435$, нет оснований отвергнуть нулевую гипотезу – партия деталей может быть принята.

Замечание 3.3. Если рассматривается гипотеза $\sigma^2 \neq \sigma_0^2$:

а) в пакете *Excel* с помощью функции ХИ2.ОБР() находятся границы двусторонней критической области $\chi^2_{\text{лев.крит}} = \text{ХИ2.ОБР}(\frac{\alpha}{2}; n-1)$; и $\chi^2_{\text{прав.крит}} = \text{ХИ2.ОБР}(1-\frac{\alpha}{2}; n-1)$. Если $\chi^2_{\text{лев.крит}} < \chi^2_{\text{набл}} < \chi^2_{\text{прав.крит}}$ – нет оснований отвергнуть нулевую гипотезу.

б) в пакете STATISTICA границы двусторонней критической области можно найти с помощью вероятностного калькулятора (см. рис.3.8)

Рис. 3.8 – расчет границ двусторонней критической области в пакете STATISTICA
1) $\chi^2_{\text{лев.крит}}$; 2) $\chi^2_{\text{прав.крит}}$

Контрольный пример 3.2. 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.

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 -значение.

Введем выборочные данные в столбцы A и B листа *MS Excel* (рис. 3.8 – 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. 3.9). 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 ОК .

5) На экране в диапазоне D4: F13 появится таблица результатов решения (рис. 3.9)

E15 =F.ТЕСТ(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.ТЕСТ	0,67966	

Рис. 3.9 – Исходные данные и результаты решения первой части контрольного примера 3.2

Здесь символом F обозначено наблюдаемое значение статистики F. Символ $P\ F \leq f$ обозначает статистическую значимость $p = P\ F\ k_1, k_2 \geq f$ – если $p \leq \alpha$, то нулевая гипотеза отвергается, а символ *F критическое одностороннее* – критическое значение $F_{\text{кр}}\ \alpha, k_1, k_2$ 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 $F_{\text{кр}}\ 0.025{,}9{,}6\ = 5{,}52341$ 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 3.4. Если бы в качестве альтернативы выступала гипотеза $\text{H}_1: \sigma_1^2 > \sigma_2^2$, в поле **Альфа** надо было ввести число 0,05. При этом процедура выдала бы следующие результаты (рис. 3.10):

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. 3.10 – Результаты решения контрольного примера 3.2 при $\text{H}_1: \sigma_1^2 > \sigma_2^2$

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.

Замечание 3.5. На рисунке 3.9 в ячейках E15 приведена проверка гипотезы о равенстве дисперсий с помощью функции **F.ТЕСТ.** В ячейке D15 приведена значимость $p = 2P\ F_{11,11} \geq 1{,}109\ = 0{,}8668$ (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.

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, т.е. в числителе формулы находится величина $x - y - \delta$.

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 3.11).

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. 3.11 и щёлкнем на кнопке *ОК*.

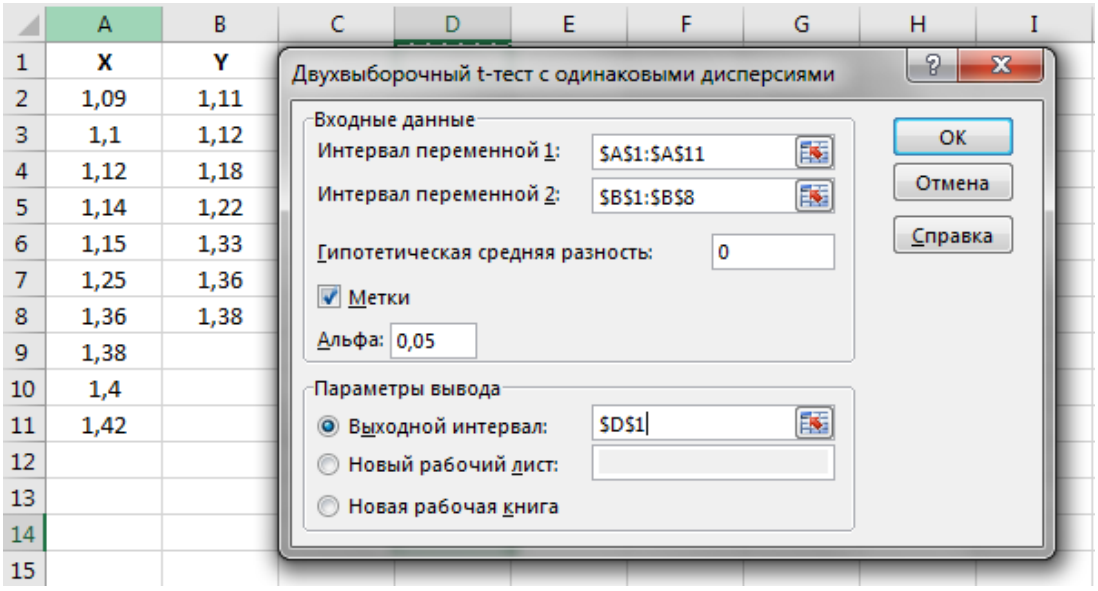


Рис. 3.11 – Диалоговое окно процедуры Двухвыборочный t-тест с одинаковыми дисперсиями

г) На экране в диапазоне D1:F14 появится таблица результатов решения (см. рис. 3.12)

	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. 3.12 – Проверка гипотезы о равенстве средних двух нормальных генеральных совокупностей с неизвестными равными дисперсиями

В таблице 3.1 приведены некоторые термины, используемые в таблице результатов, и дано их краткое объяснение.

Таблица 3.1

Заголовок	**Объяснение**
Объединённая дисперсия	Выборочная дисперсия $s^2 = \frac{n_1-1\ s_x^2 + n_2-1\ s_y^2}{n_1+n_2-2}$, вычисленная по объединённым данным обеих выборок
df	Число степеней свободы статистики T: $df = n_1 + n_2 - 2$
t-статистика	Расчётное значение статистики T, найденное по формуле (3.3)
P(T<=t) одностороннее	Значимость p. При альтернативе H_1: $a_1 - a_2 > \delta \ - \ p = P\ \mathrm{T}_{n_1+n_2-2} \geq t$ При альтернативе H_1: $a_1 - a_2 < \delta \ - \ p = P\ \mathrm{T}_{n_1+n_2-2} \leq t$
t критическое одностороннее	Критическое значение $t_{кр}\ \alpha,\ df$ порядка α распределения Стьюдента с $df = n_1 + n_2 - 2$ степенями свободы.
P(T<=t) двухстороннее	Значимость p при альтернативе H_1: $a_1 - a_2 \neq \delta$
t критическое двухстороннее	Критическое значение $t_{кр}\ \frac{\alpha}{2},\ df$ порядка $\frac{\alpha}{2}$ распределения Стьюдента с $df = n_1 + n_2 - 2$степенями свободы.

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*.

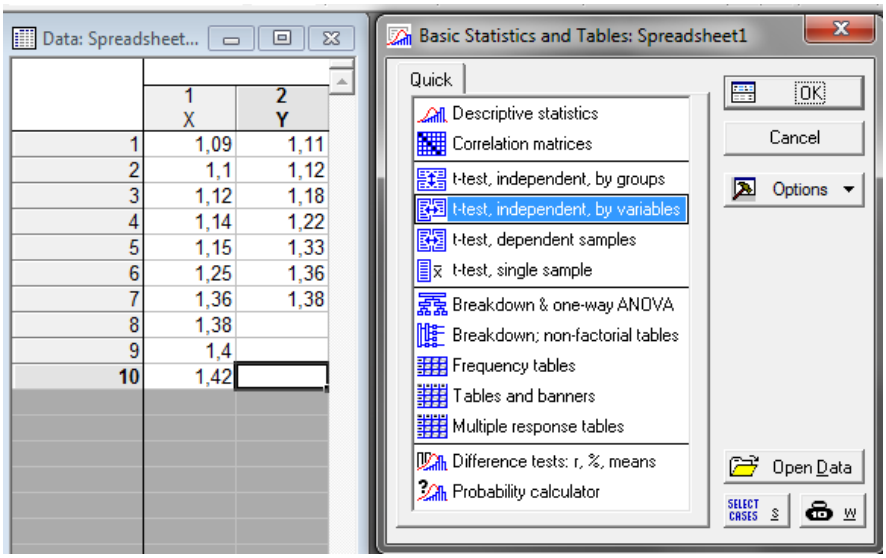


Рис. 3.13 – Исходные данные

Выбираем *t-test, independent, by variables*. Далее зададим диапазон исходных данных, нажав на кнопку *Variables* и выбрав там: *X* и *Y*. Далее нажмём кнопку *OK* (рис. 3.14).

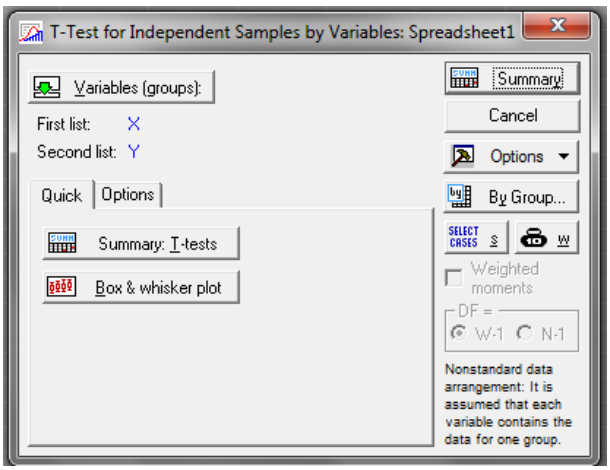
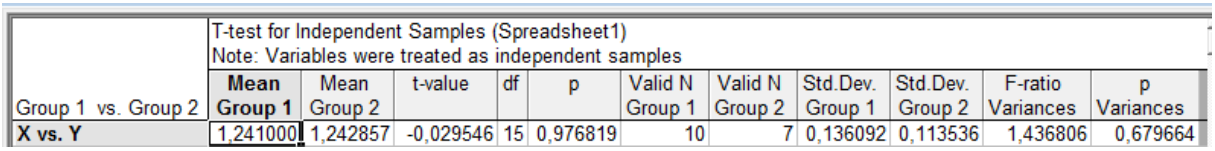


Рис. 3.14 – Задание диапазона исходных данных

Нажав кнопку *Summary*, получаем таблицу (рис.3.15):



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. 3.15 – 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 $\mathrm{H}_1: a_1 \neq a_2$.

Замечание 3.6. 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 $p\,Variances < \alpha$), следует воспользоваться другой формулой для проверки гипотезы о равенстве средних (см. формулы (3.4) – (3.6)), либо воспользоваться непараметрическими критериями сравнения двух выборок.

В данном примере гипотеза о равенстве дисперсий принимается, так как $p\,Variances = 0{,}679664 > 0{,}05$.

Контрольный пример 3.3. 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.

Принять в качестве альтернативной гипотезу $\mathrm{H}_1: a_1 \neq a_2$.

Решение. В пакете *Excel* проверяется гипотеза $\mathrm{H}_0: a_1 - a_2 = \delta$ 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.*

На рис. 3.16 изображено диалоговое окно этой процедуры. Она полностью идентично диалоговому окну процедуры *Двухвыборочный t-тест с одинаковыми дисперсиями.*

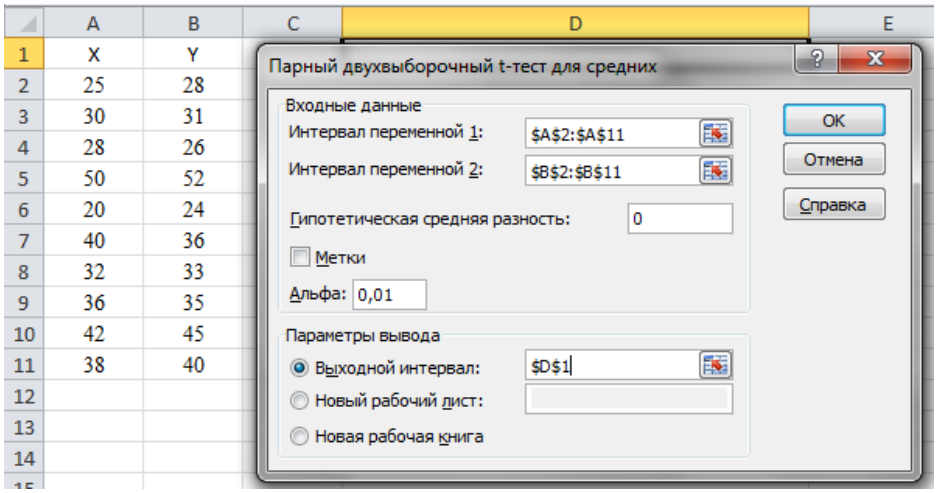


Рис. 3.16 – Исходные данные задачи и диалоговое окно процедуры *Парный двухвыборочный t-тест для средних*

Замечание 3.7. В пакете *Excel* проверка гипотезы $H_0: a_1 - a_2 = \delta$ с помощью процедуры *Парный двухвыборочный t-тест для средних* осуществляется следующим образом:

а) при альтернативе $H_1: a_1 - a_2 > \delta$ строится правосторонняя критическая область. Областью отклонения гипотезы является интервал $[t_{кр}\ \alpha, n-1\ ; \infty)$; $t_{кр}\ \alpha, n-1$ – критическое значение порядка α распределения Стьюдента с $k = n - 1$ степенями свободы.

б) при альтернативе $H_1: a_1 - a_2 < \delta$ строится левосторонняя критическая область. Область отклонения гипотезы – интервал $(-\infty; -t_{кр}\ \alpha, n-1\ ]$.

в) при альтернативе $H_1: a_1 - a_2 \neq \delta$ строится двухсторонняя критическая область, для которой областью принятия гипотезы является интервал $-t_{кр}\ \frac{\alpha}{2}, n-1\ ;\ t_{кр}\ \frac{\alpha}{2}, n-1\ \ .$

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. 3.17 – Проверка гипотезы о равенстве средних результатов взвешиваний

Результаты проверки гипотезы приведены на рис. 3.17 в диапазоне ячеек D1: F14.

Выборочный коэффициент корреляции $r_{\text{В}} = 0{,}959$ (ячейка E7) свидетельствует о высокой корреляционной зависимости результатов двух серий измерений. Это обстоятельство подтверждает правильность выбора критерия проверки нулевой гипотезы.

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 $p > \alpha$.

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.3.18).

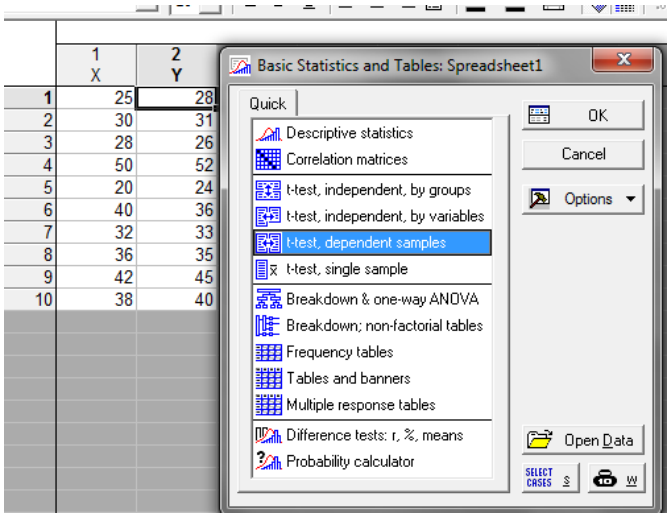


Рис. 3.18 – Выбор *t-test, dependent samples*

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 (рис. 3.19):

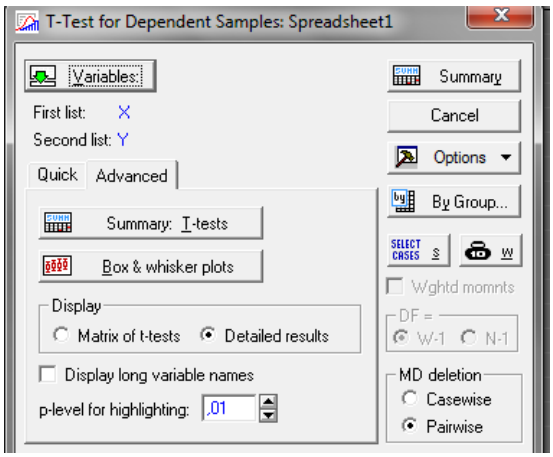


Рис. 3.19 – Вкладка *Advanced* окна *t-test for dependent samples*

Нажав кнопку *Summary*, получаем таблицу (рис. 3.20).

T-test for Dependent Samples (Spreadsheet1)
Marked differences are significant at p < ,01000

Variable	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. 3.20 – Проверка гипотезы о равенстве средних результатов взвешиваний

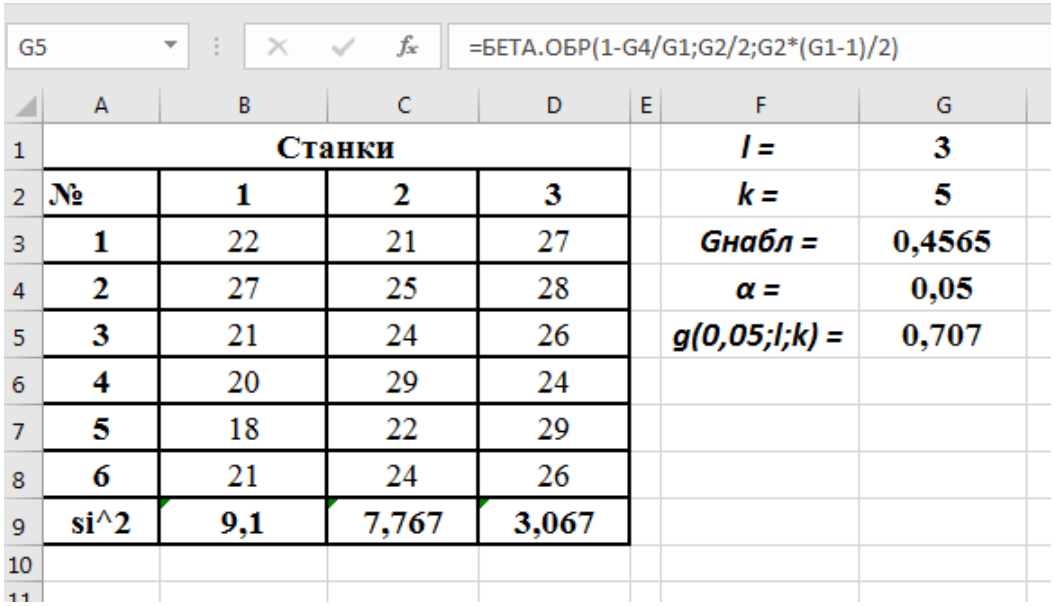
Из заголовка таблицы следует, что для наличия значимых различий статистическая значимость (p-значение) должна быть меньше 0,01 (в нашем случае $p = 0{,}286933$). Значит, гипотеза о равенстве средних принимается – что подтверждает вывод, сделанный в пакете *Excel*.

Контрольный пример 3.4. С трех станков, настроенных на обработку одних и тех же деталей, было отобрано по четыре образца. С помощью критерия Кохрена определить, одинаковая ли точность станков. Принять $\alpha = 0{,}05$.

Номер станка	Результаты измерений					
	1	2	3	4	5	6
1	22	27	21	20	18	21
2	21	25	24	29	22	24
3	27	28	26	24	29	26

Решение. 1) Точность можно оценить с помощью дисперсии размеров деталей, обработанных на каждом станке. Проверяется гипотеза о равенстве дисперсий. Если дисперсии однородны, то точность станков одинакова.

На рис. 3.21 приведены результаты проверки гипотезы $D\ X_1 = D\ X_2 = \cdots = D\ X_l$ о равенстве дисперсий трех наборов данных, с помощью критерия Кохрена (в пакете Excel).



G5 | =БЕТА.ОБР(1-G4/G1;G2/2;G2*(G1-1)/2)

	A	B	C	D	E	F	G
1	Станки					*l* =	3
2	№	1	2	3		*k* =	5
3	1	22	21	27		*Gнабл* =	0,4565
4	2	27	25	28		*α* =	0,05
5	3	21	24	26		*g(0,05;l;k)* =	0,707
6	4	20	29	24			
7	5	18	22	29			
8	6	21	24	26			
9	si^2	9,1	7,767	3,067			
10							

Рис. 3.21 – Проверка гипотезы о равенстве нескольких дисперсий с помощью критерия Кохрена

В диапазоне B9: K9 находятся выборочные дисперсии трех «наборов» данных, характеризующих точность станков. Дисперсии вычислены с помощью формулы ДИСП. В B3: B8 , введённой первоначально в ячейку B9 и скопированной затем в ячейки C9: K9.

В ячейке G3 находится выборочное значение статистики G, найденное с помощью формулы = МАКС(B9: K9)/СУММ(B9: K9), а в ячейке N5 – критическое значение G_{kp} 0,05; 3; 5 = 0,707 этой статистики, вычисленное с помощью встроенной функции БЕТА. ОБР .

Так как $G_{\text{набл}} < G_{kp}$, то гипотеза о равенстве дисперсий принимается – станки обеспечивают одинаковую точность.

2) Введем исходные данные в созданную таблицу в пакете STATISTICA, как показано на рисунке 3.22:

	1 Var1	2 Var2			
1	1	22	10	2	29
2	1	27	11	2	22
3	1	21	12	2	24
4	1	20	13	3	27
5	1	18	14	3	28
6	1	21	15	3	26
7	2	21	16	3	24
8	2	25	17	3	29
9	2	24	18	3	26

Рис. 3.22 – Исходные данные

Выполним команду *Statistics/ANOVA*. Появится меню *General ANOVA/MANOVA*. Выполним установки, как показано на рисунке 3.23.

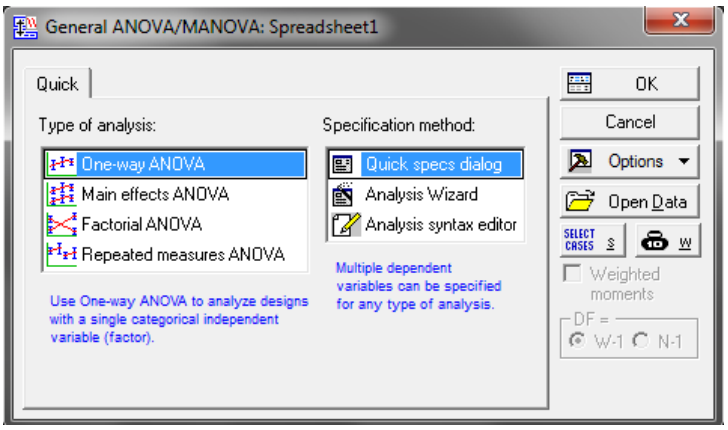


Рис. 3.23 – Стартовая панель модуля *ANOVA*

После нажатия кнопки *OK* в появившемся окне выберем переменные для анализа (с помощью кнопки *Variables*). После того как кнопка будет нажата, на экране появится диалоговое окно *Select dependent variables and categorical predictor (factor)* (Выбрать списки зависимых переменных и факторов).

В левой части окна имя переменной выберем зависимую (*depended*) переменную (*VAR2*), а в правой – фактор (*VAR1*). После нажатия кнопки *OK* в появившемся окне выберем *All*, нажав кнопку *Factor codes* (рисунок 3.24).

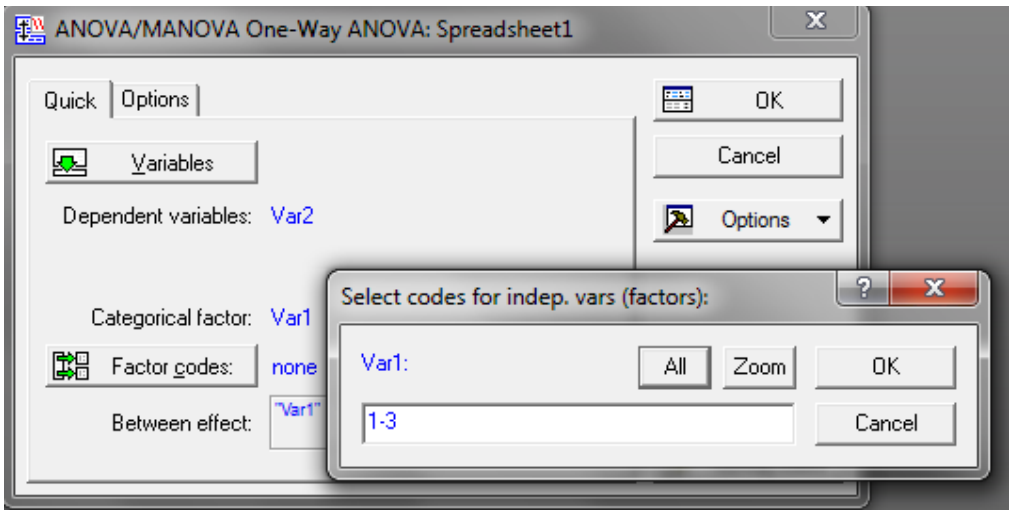


Рис. 3.24 Окно выбора факторов

Нажмем кнопку *OK* в правом углу стартовой панели. На экране появится диалоговое окно *Anova Results 1*.

В левом нижнем углу этого диалогового окна нажмем клавишу *More results* (Больше), перейдя, таким образом, к развернутому представлению результатов.

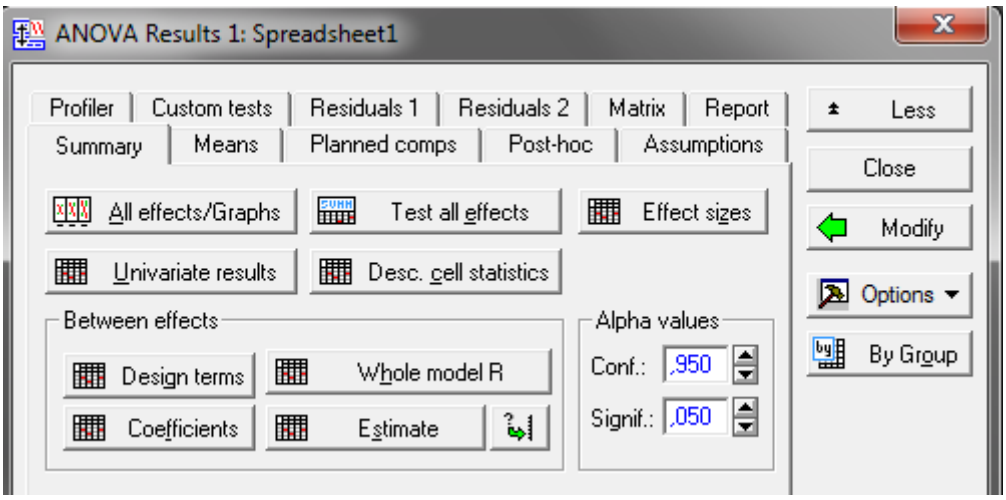


Рис.3.25 – Развернутое представление результатов

Для проверки дисперсий на однородность перейдем к вкладке *Assumptions* (Предположения). Нажмем на кнопку, где указан тест Кохрена, Хартли, Бартлетта (Cohran C, Hartley, Bartlett). Появится следующая таблица:

Tests of Homogeneity of Variances (Spreadsheet1)
Effect: "Var1"

	Hartley F-max	Cochran C	Bartlett Chi-Sqr.	df	p
Var2	2,967391	0,456522	1,389658	2	0,499160

Рис. .17 Проверка дисперсий на однородность

Как видно из таблицы, проверка дисперсий на однородность осуществляется одновременно по 3 тестам.

Так как статистическая значимость больше 0,05, то принимается нулевая гипотеза – точность станков одинакова.

3.3. Задания для самостоятельной работы.

Задание 1. Проектный, контролируемый размер изделий, изготовляемых станком автоматом $a = a_0$ мм. Измерения 20 случайно отобранных изделий дали результаты, приведённые в таблице. Требуется:

а) при уровне значимости 0,05 проверить нулевую гипотезу $\mathrm{H}_0: a = 35$ (станок обеспечивает проектный размер деталей) при конкурирующей гипотезе $\mathrm{H}_1: a \neq 35$;

б) партия деталей принимается, если дисперсия контролируемого размера значимо не превышает $\sigma_0^2 = 0{,}2$. Можно ли принять партию при уровне значимости 1) 0,01; 2) 0,05? Принять в качестве альтернативной $\mathrm{H}_1: \sigma^2 > \sigma_0^2$.

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	30,0	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	25,0	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	35,0	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	36,0	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	26,0
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

Задание 2.

Вариант 1. Испытывались на растяжение образцы, у которых обработка поверхностей производилась двумя различными методами. Результаты испытаний приведены в таблице:

Метод 1	16	14	19	20	15	18	18	19	17	18
Метод 2	13	19	14	14	15	10	17	21	13	15

Требуется решить при уровне значимости 0.05, могут ли данные испытаний принадлежать нормально распределённым совокупностям с одинаковыми средними значениями. Необходимо проверить гипотезы: $H_0: a_1 = a_2$; $H_1: a_1 \neq a_2$ Предварительно проверить гипотезу о равенстве дисперсий. Предполагается, что случайные величины X и Y распределены нормально.

Вариант 2. Группа школьников в течение летних каникул находилась в спортивном лагере. До и после сезона у них измерили жизненную емкость легких.

До «эксперимента» (x_i, мл):	3400	3600	3000	3500	2900	3100	3200	3400	3200	3400
После «эксперимента» (y_i, мл):	3800	3700	3300	3600	3100	3200	3200	3300	3500	3600

По результатам измерений нужно определить, значимо ли изменился этот показатель под влиянием интенсивных физических упражнений. Принять в качестве конкурирующей гипотезу $H_0: a_1 \neq a_2$ и $\alpha = 0{,}01$. Предварительно проверить гипотезу о равенстве дисперсий.

Вариант 3. Для сравнения качества однотипных батарей электропитания, выпускаемых двумя предприятиями, были испытаны 2 контрольные группы батарей по 10 батарей в каждой. В ходе испытаний фиксировалось время (в часах) каждой батареи при работе на стандартную нагрузку. Результаты испытаний приведены в следующей таблице:

Батарея А	191	178	278	199	156	228	147	266	220	159
Батарея В	190	227	182	161	212	194	173	208	196	215

Проверить на уровне значимости 0.05 гипотезу $H_0: a_1 = a_2$ о том, что средние времена разрядки батарей, выпускаемых обоими предприятиями, одинаковы. Альтернативная гипотеза – $H_1: a_1 \neq a_2$. Предварительно прове-

рить гипотезу о равенстве дисперсий, приняв в качестве альтернативной – $H_1: \sigma_1^2 > \sigma_2^2$.

Вариант 4. Производительность двух моторных заводов, выпускающих дизельные двигатели, характеризуется следующими данными:

1-й завод	72	84	69	74	82	67	75	86	68	61
2-й завод	55	65	73	66	58	71	77	68	68	59

Можно ли считать одинаковыми производительность дизельных двигателей на обоих заводах при уровне значимости 0.05? Альтернативная гипотеза – $H_1: a_1 \neq a_2$. Предварительно проверить гипотезу о равенстве дисперсий.

Вариант 5. Доходы аптек одного из микрорайонов города за некоторый период составили **128; 192; 223; 398; 205; 266; 219; 260; 264; 98** (условных единиц). В соседнем микрорайоне за то же время они были равны **286; 240; 263; 266; 484; 223; 335.** Предполагая, что данная случайная величина имеет нормальное распределение, проверить по критерию Фишера гипотезу о равенстве генеральных дисперсий. По критерию Стьюдента проверить гипотезу о равенстве генеральных средних (альтернативная гипотеза – об их неравенстве). Во всех расчётах уровень значимости α = 0,05.

Вариант 6. На станке-автомате обрабатываются втулки. Взяты 2 пробы, по 10 штук каждая:

x_i	2,42	2,5	1,44	1,94	1,86	2,05	2,21	1,96	2,29	2,31
y_i	2,34	1,66	2,17	1,89	1,76	2,21	2,12	1,88	2,25	2,1

Распределение диаметров втулок – нормальное. Проверить гипотезу о том, что режим работы станка не изменится, если генеральные средние в момент выбора проб – разные. Альтернативная гипотеза – $H_1: a_1 \neq a_2$. Принять $\alpha = 0{,}05$.

Предварительно проверить гипотезу о равенстве дисперсий.

Вариант 7. Приведены результаты тестирования в экспериментальной и

контрольной группах, по 9 студентов в каждой. Есть ли статистически значимые различия по среднему значению признака?

Экспериментальная группа	48	36	28	46	36	24	50	38	26
Контрольная группа	39	21	44	31	26	36	24	16	20

Принять в качестве конкурирующей гипотезу $H_1: a_1 \neq a_2$

Предварительно проверить гипотезу о равенстве дисперсий. Принять $\alpha = 0.05$.

Вариант 8. Из первого нарезного оружия было произведено 8 выстрелов. При этом измерялись начальные скорости пуль. Получены следующие результаты:

902,4; 901,3; 898,4; 903,5; 901,1; 900,4; 899,7; 900,3 (м/с).

Из второго оружия было произведено 7 выстрелов. Скорости вылета пуль оказались равны:

905,5; 910,3; 903,8; 902,4; 899,9; 903,3; 905,6 (м/с).

Предполагая, что данная случайная величина имеет нормальное распределение, по критерию Фишера проверить гипотезу о равенстве генеральных дисперсий. По критерию Стьюдента проверить гипотезу о равенстве генеральных средних (альтернативная гипотеза – об их неравенстве). Принять $\alpha = 0{,}05$.

Задание 3. Требуется при уровне значимости α проверить гипотезу о равенстве математических ожиданий двух коррелированных (зависимых) нормально распределённых генеральных совокупностей по извлечённым из них выборкам. Задачу решить с применением пакетов *Statistica* и *Excel*. Принять в качестве альтернативной гипотезу $H_1: a_1 \neq a_2$.

Вариант 1 $\alpha = 0{,}05$										
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	
Вариант 2 $\alpha = 0{,}01$										
X	16	14	14	23	11	12	17	14	18	16
Y	13	10	11	21	6	9	16	10	16	13

Вариант 3 $\alpha = 0{,}01$										
X	2.5	2.1	5.1	2.34	7.63	1.9	1.12	1.49	3.58	3.4
Y	2.6	2.6	5.63	2.38	7.61	2.8	1.12	1.45	3.58	3.5
Вариант 4 $\alpha = 0{,}01$										
X	3	8	4	4	7	8	2	5	6	3
Y	4	5	2	5	6	8	3	4	5	5
Вариант 5 $\alpha = 0{,}05$										
X	1	2	5	10	14	14	15	16	17	19
Y	3	4	6	7	9	12	13	15	18	20
Вариант 6 $\alpha = 0{,}05$										
X	3.5	3.1	5.1	3.34	3.63	3.9	5.12	3.49	3.58	3.4
Y	4.6	4.6	5.63	4.38	4.61	4.8	4.12	4.45	3.58	3.8
Вариант 7 $\alpha = 0{,}01$										
X	6	10	8	20	22	25	20	32	35	38
Y	5	18	6	27	14	10	18	35	28	30
Вариант 8 $\alpha = 0{,}05$										
X	76	71	57	49	70	69	26	65	59	
Y	81	85	52	52	70	63	33	83	62	

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 $\alpha = 0{,}05$.

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

Лабораторная работа 4

ДИСПЕРСИОННЫЙ АНАЛИЗ

Цель работы: изучить методики применения дисперсионного анализа при проверке гипотезы о равенстве математических ожиданий, либо при установлении того, оказывает ли качественный фактор F существенное влияние на исследуемую величину X.

Используемые программные средства: MS Excel 2010 (2016), STATISTICA 8.0.

4.1. Краткие теоретические сведения.

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 $X_1, X_2, \dots, X_m$, где m – число уровней фактора. Каждая случайная величина X_j соответствует определённому уровню фактора F_j и для неё получена выборка значений $x_{1j}, x_{2j}, \dots, x_{nj}$, где n_j – число наблюдений для данного уровня. Данные наблюдений можно представить в виде таблицы 4.1, в которой количество элементов в столбце может быть различным. При этом $n = n_1 + n_2 + \cdots + n_m$ – общее число всех наблюдений.

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

$$H_0: M\ X_1\ = M\ X_2\ = \cdots = M\ X_m\ .$$

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

Таблица 4.1

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 x_j	x_1	x_2		x_m

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

Дисперсионный анализ может быть применён, если:

1. генеральные совокупности $X_1, X_2, \ldots, X_m$ 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:

$x_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij} \quad j = 1, m$ – *групповые средние*

$x = \sum_{j=1}^{m} \sum_{i=1}^{n_j} n_{ij}$ – *общая выборочная средняя.*

Несмещённой оценкой для неизвестной дисперсии σ^2 является сумма квадратов $\sum_{j=1}^{m} \sum_{i=1}^{n_j} x_{ij} - x^{\ 2}$, делённая на $n - 1$, где $n = \sum_{j=1}^{m} n_j$ – количество всех наблюдений (если на каждом уровне проведено одинаковое количество наблюдений $n_1 = n_2 = \cdots = n_m = n'$, 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 x_j.

Обозначим

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

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

$Q_{\text{ост}} = \sum_{j=1}^{m} \sum_{i=1}^{n_j} x_{ij} - x_j^{\,2}$ – *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:

$$Q_{\text{общ}} = Q_{\text{факт}} + Q_{\text{ост}} \tag{4.1}$$

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 $s^2_{\text{факт}} = \frac{Q_{\text{факт}}}{m-1}$, так как при расчёте $Q_{\text{факт}}$ используется m групповых средних, связанных между собой одним уравнением. Несмещённой оценкой внутригрупповой (остаточной) дисперсии является $s^2_{\text{ост}} = \frac{Q_{\text{ост}}}{n-m}$, ибо при расчёте $Q_{\text{ост}}$ используются все n наблюдений, связанных между собой m уравнениями $x_j = \frac{1}{n_j} \sum_{i=1}^{n_j} x_{ij}$. 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):

$$F = \frac{s^2_{\text{факт}}}{s^2_{\text{ост}}} \tag{4.2}$$

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 $F_{kp}\ \alpha, k_1, k_2$ и принимается, если $F < F_{kp}\ \alpha, k_1, k_2$.

Степень влияния фактора на результативный показатель может быть измерена с помощью выборочного коэффициента детерминации

$$R^2 = \frac{Q_{\text{факт}}}{Q_{\text{ост}}}$$

показывающего, какова доля общей вариации объясняется влиянием исследуемого фактора.

Замечание 4.1. Если факторная и остаточная дисперсии различаются незначимо, то влияние фактора можно считать незначительным и, следовательно, принять гипотезу о равенстве математических ожиданий.

Замечание 4.2. Границу правосторонней критической области можно найти, используя:

- в пакете *Excel 2010* – функцию $\text{F. ОБР. ПХ}(\alpha, k1, k2)$;
- таблицу критических точек распределения Фишера-Снедекора.

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

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

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

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

$Q_e = \sum_{i=1}^{a} \sum_{j=1}^{b} x_{ij} - x_i - x_j + x^{\ 2}$ – необъяснённая сумма квадратов отклонений или сумма квадратов отклонений ошибки;

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

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

$x_i = \frac{1}{a} \sum_{j=1}^{a} x_{ij}$ – 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,

где

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

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

$k_e = a - 1 \cdot b - 1$ – 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:

$H_0: m_{1A} = m_{2A} = \cdots = m_{aA}$; H_1: не все m_{iA} равны;

для фактора B:

$H_0: m_{1B} = m_{2B} = \cdots = m_{aB}$; H_1: не все m_{iB} равны.

Чтобы определить влияние фактора A нужно наблюдаемое отношение Фишера $F_a = \frac{s_a^2}{s_e^2}$ сравнить с критическим отношением Фишера $F_{kp}\ \alpha;\ k_A, k_e$.

Чтобы определить влияние фактора B, нужно фактическое отношение Фишера $F_a = \frac{s_b^2}{s_e^2}$ сравнить с критическим отношением Фишера $F_{kp}\ \alpha;\ k_B, k_e$.

Если фактическое отношение Фишера больше критического отношения Фишера, то следует отклонить нулевую гипотезу с уровнем значимо-

сти 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 – А и В, но и возможное взаимодействие факторов А и В. Тогда a – число групп фактора А и b – число групп фактора В, r – число повторений; n – число наблюдений в каждой группе

В таблице 4.2 приведена схема двухфакторного дисперсионного анализа с повторениями:

Таблица 4.2

	Сумма квадратов	**Число степеней свободы**	**Дисперсия**	**Критерий Фишера**
Фактор А	Q_A	$k_A = a - 1$	$s_A^2 = \frac{Q_A}{k_A}$	$F_A = \frac{s_A^2}{s_e^2}$
Фактор В	Q_B	$k_B = b - 1$	$s_B^2 = \frac{Q_B}{k_B}$	$F_B = \frac{s_B^2}{s_e^2}$
Взаимодействие АхВ	Q_{AB}	$k_{AB} = a - 1 \quad b - 1$	$s_{AB}^2 = \frac{Q_{AB}}{k_{AB}}$	$F_{AB} = \frac{s_{AB}^2}{s_e^2}$
Ошибка	Q_e	$k_e = ab \ n - 1$	$s_e^2 = \frac{Q_e}{k_e}$	

Здесь:

$Q_A = \frac{a}{N} \ \sum_{i=1}^{a} c_i^2 - \frac{c^2}{N} \qquad c = \quad x_{ijk};\ c_i^2 = \ \sum_{j=1}^{b} \ \sum_{k=1}^{r} x_{ijk}\ ; N = a \cdot b \cdot n$ – сумма квадратов для фактора А;

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

$Q_{AB} = \frac{ab}{N} \ \sum_{i=1}^{a} \ \sum_{j=1}^{b} c_{ij}^2 - Q_A - Q_B - \frac{c^2}{N} \ \ c_{ij} = \ \sum_{k=1}^{r} x_{ijk}$ – сумма квадратов для взаимодействия А × В;

$Q_A = c_0 - \frac{1}{n} \ \sum_{i=1}^{a} \ \sum_{j=1}^{b} c_{ij}^2 \ \ c_0 = \ \ x_{ijk}^2$ – 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.

4.2. Практическая часть.

Контрольный пример 4.1. Исследуется зависимость процентного содержания брака (величина X) среди изделий, изготовленных за единицу времени, от температуры окружающей среды (фактор A). Был произведён подсчёт количества бракованных изделий для пяти интервалов времени при трёх различных температурах окружающей среды. Результаты измерений представлены в следующей таблице.

F_i	x_{ij}				
Процент брака при повышенной температуре	2,5	3,3	2,4	3	2,6
Процент брака при нормативной температуре	2,4	3,2	2,2	2,7	2,3
Процент брака при пониженной температуре	2,6	3,4	3	3,1	2,8

Методом дисперсионного анализа проверить гипотезу о влиянии температуры среды на процентное содержание брака среди изготовленных изделий и определить степень этого влияния.

Решение. Пусть уровень фактора F_1 соответствует повышенной температуре, F_2 – нормативной температуре, F_3 – пониженной температуре.

Введем исходные данные на лист MS Excel, как показано на рис. 4.1.

	A	B	C
1	**% брака при повышенной температуре**	**% брака при нормативной температуре**	**% брака при пониженной температуре**
2	2,5	2,4	2,6
3	3,3	3,2	3,4
4	2,4	2,2	3
5	3	2,7	3,1
6	2,6	2,3	2,8
7	**Групповые средние**		
8	**2,76**	**2,56**	**2,98**
9	**Общая средняя**		
10		**2,7667**	
11	**Групповые суммы квадратов отклонений**		
12	**0,572**	**0,652**	**0,368**
13	**Общая сумма квадратов отклонений**		
14		**2,0333**	
15	**Общее число наблюдений**	**15**	
16	**Число уровней фактора**	**3**	

Рис. 4.1 – Вид листа MS Excel с исходными данными и расчетами для дисперсионного анализа

Для каждого фактора рассчитаем групповую среднюю. Для этого в ячейку A8 введем формулу = СРЗНАЧ(A2:A6), которую затем скопируем методом автозаполнения вправо по строке. Общую среднюю рассчитаем с помощью функции = СРЗНАЧ(A2:C6). Общее число наблюдений рассчитаем с помощью функции СЧЕТ A2:C6 , которая подсчитывает число заполненных числами ячеек в заданном диапазоне, а пустые ячейки игнорирует (рис. 4.1).

Результаты расчетов по дисперсионному анализу с помощью функций MS Excel оформим так, как показано на рис. 4.2

D	E	F	G	H	I	J
	Сумма квадратов отклонений	**Число степеней свободы**	**Дисперсия**	**Fнабл**	**p-значение**	**Fкрит**
межгрупповая	0,4413	2	0,2207	1,6633	0,2304	3,8853
внутригрупповая	1,592	12	0,1327			
общая	2,0333					
выборочный коэффициент детерминации			**0,2170**			

Рис. 4.2 – результаты решения исходной задачи с помощью функций пакета MS Excel

Общая сумма квадратов отклонений уже рассчитана, поэтому в ячейке E4 поставим ссылку на ячейку B14. Внутригрупповую (остаточную) сумму квадратов отклонений рассчитаем в ячейке E3, сложив соответствующие значения для всех факторов = СУММ A12:C12 .

Межгрупповую (факторную) сумму квадратов отклонений найдем как разность значений в ячейках E4 и E3.

Введем в ячейки F2 и F3 число степеней свободы: для межгрупповой дисперсии это $k_1 = m - 1 = 3 - 1 = 2$, для внутригрупповой дисперсии $k_2 = n - m = 15 - 3 = 12$.

Рассчитаем межгрупповую и внутригрупповую дисперсии в ячейках G2 и G3, разделив соответствующие суммы квадратов отклонений на число степеней свободы.

Наблюдаемое значение критерия Фишера вычислим, разделив межгрупповую (факторную) дисперсию на внутригрупповую (остаточную) дисперсию (ячейка H2). Таким образом, $F_{\text{набл}} \approx 1{,}66$.

Для расчета критической точки распределения Фишера в ячейке J2 используем функцию Excel F.ОБР.ПХ(0,05; F2; F3). Так как $F_{\text{набл}}$ меньше критического значения $F_{\text{крит}} = 3{,}88$, то принимается гипотеза H_0. Результаты анализа показывают, что влияние температурного режима на процентное содержание брака в готовой продукции не является существенным.

Выборочный коэффициент детерминации:

$$R^2 = \frac{Q_{\text{факт}}}{Q} = \frac{0{,}4413}{2{,}0333} \approx 0{,}22$$

рассчитан в ячейке G5. Он означает, что только 22% брака в готовой продукции *обусловлены* температурным режимом.

Аналогичные результаты получим с помощью инструмента из *Пакета анализа*.

Из раздела меню *Данные/Анализ данных* выберем *Однофакторный дисперсионный анализ*. Заполним диалоговое окно, как показано на рис. 4.3.

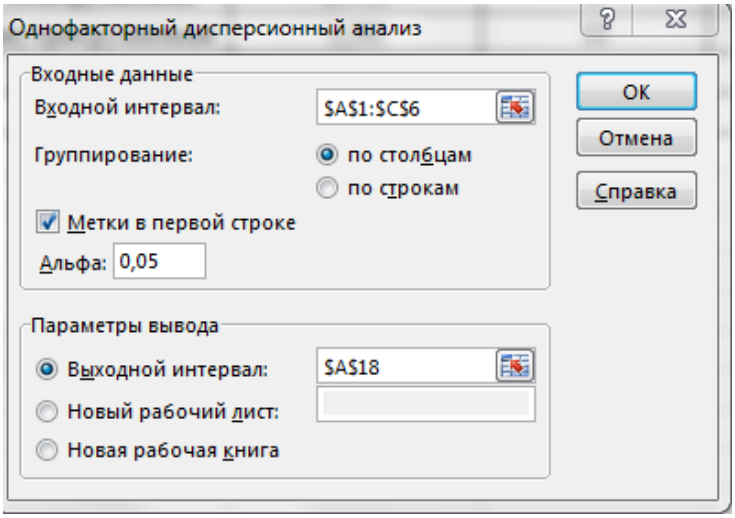


Рис. 4.3 – Диалоговое окно *Однофакторный дисперсионный анализ*

Флажок *Метки в первой строке* поставлен потому, что входной интервал включает заголовки столбцов, и они будут использованы для формирования результата.

В результате действия процедуры выводятся две таблицы. Первая таблица – *Итоги* (см. рис. 4.3). В ней показаны выборочные характеристики для каждого уровня фактора: количество наблюдений (счёт), сумма значений, среднее и дисперсия.

18	Однофакторный дисперсионный анализ				
19					
20	ИТОГИ				
21	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
22	% брака при повышенной температуре	5	13,8	2,76	0,143
23	% брака при нормативной температуре	5	12,8	2,56	0,163
24	% брака при пониженной температуре	5	14,9	2,98	0,092

Рис. 4.4. – Таблица *Итоги*

Во второй таблице – *Дисперсионный анализ* (см. рис. 4.5) – содержатся данные о величинах для фактора между группами и внутри групп и итоговых. Это сумма квадратов отклонений (SS), число степеней свободы (df), дисперсия (MS). В последних трёх столбцах - фактическое значение отношения Фишера(F), p-уровень (P-Значение) и критическое значение отношения Фишера (F критическое).

Эти результаты аналогичны тем, что были ранее рассчитаны с помощью стандартных функций MS Excel.

27	**Дисперсионный анализ**						
28	Источник вариации	SS	df	MS	F	P-Значение	F критическое
29	Между группами	0,4413	2	0,2207	1,6633	0,2304	3,8853
30	Внутри групп	1,592	12	0,1327			
31							
32	**Итого**	**2,0333**	14				

Рис. 4.5. – Таблица *Дисперсионный анализ*

Решение в пакете STATISTICA.

Введем исходные данные в созданную таблицу в формате пакета *Statistica*, как показано на рисунке 4.6:

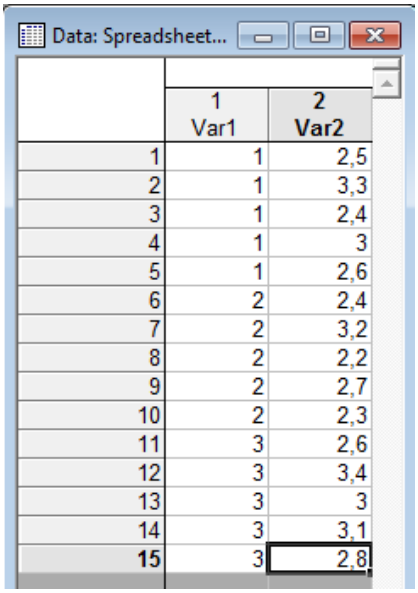


Рис 4.6 – Исходные данные Var1 – факторы; Var2 – независимая переменная.

Выполним команду *Statistics/ANOVA*. Появится меню *General ANOVA/MANOVA:*

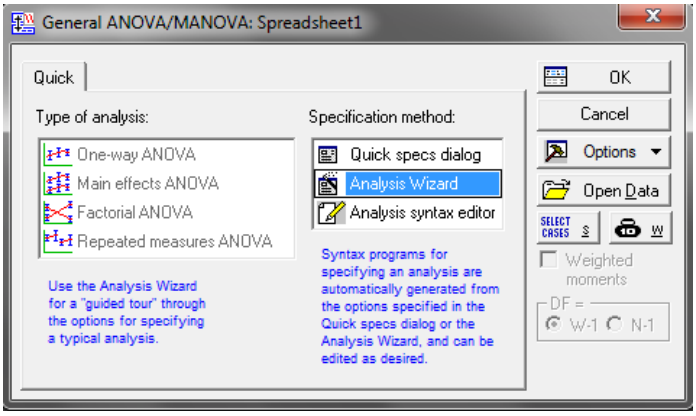


Рис. 4.7 – Стартовая панель модуля *ANOVA*

Выберем пункт *Analysis Wizard* в колонке *Specification Method* и

нажмем *OK*. Откроется окно *Variables* (рис. 4.8).

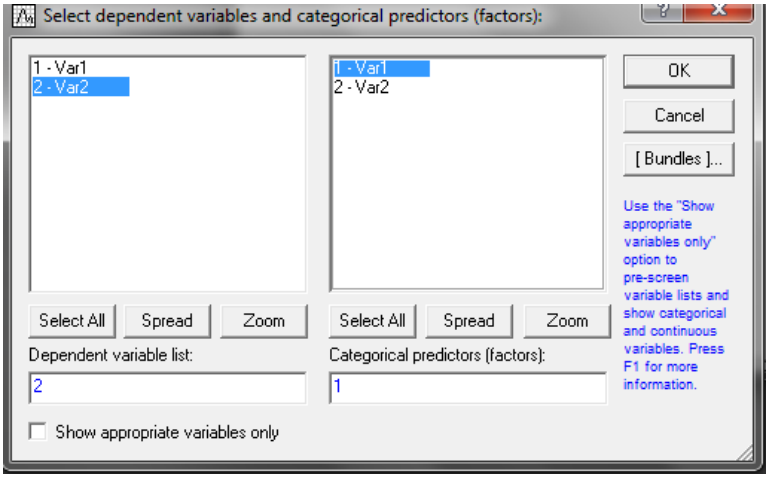


Рис. 4.8 – Окно выбора переменных для анализа

Определим независимую (VAR1) и зависимую (VAR2) переменные и нажмем ОК. Затем еще раз ОК.

Появится панель *ANOVA Results*.

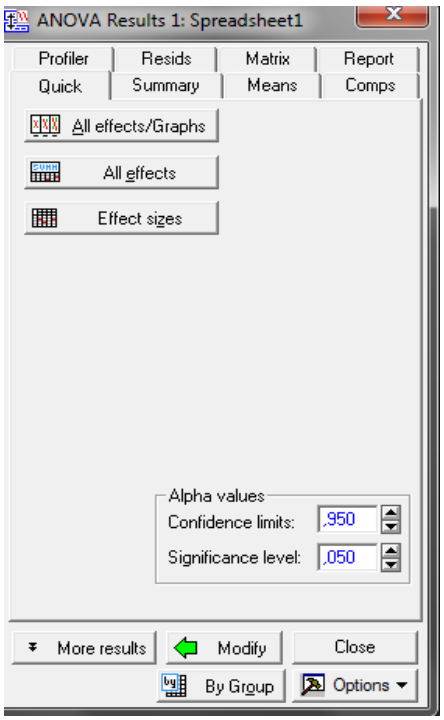


Рис. 4.9 – Диалоговое окно результатов

Для решения данной задачи достаточно нажать кнопку *All effects/Graphs*, и в открывшемся окно поставить галочку возле *SpreadSheet* и нажать *OK:*

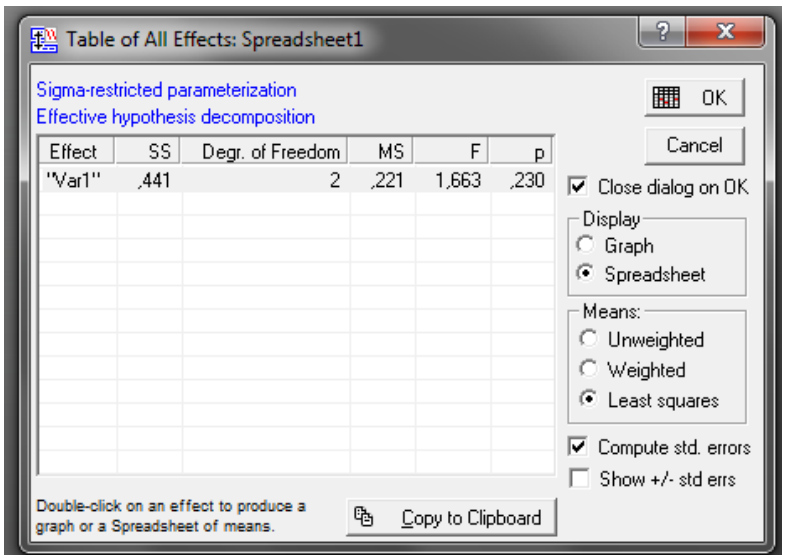


Рис. 4.10 – Диалоговое окно Table of All Effects

В окне результатов рис. 4.11 представлены результаты дисперсионного анализа:

"Var1"; LS Means (Spreadsheet1)
Current effect: F(2, 12)=1,6633, p=,23036 **1**
Effective hypothesis decomposition

Cell No.	Var1	Var2 Mean	Var2 Std.Err.	Var2 -95,00%	Var2 +95,00%	N
1	1	2,760000	0,162891	2,405092	3,114908	5
2	2	2,560000	0,162891	2,205092	2,914908	5
3	3	2,980000	0,162891	2,625092	3,334908	5

Рис. 4.11 – Результаты дисперсионного анализа

Итак, подчеркнутое предложение **1** – это ключ-решение: тут показан критерий Фишера-Снедекора, полученный в ходе решения задачи. Этот критерий надо сравнить с табличным F_{kp} Фишера-Снедекора.

Так как $F_{\text{набл}} = 1{,}633 < F_{kp} = 3{,}8853$, а также статистическая значимость $p = 0{,}23036 > \alpha = 0{,}05$ – нет оснований отвергнуть нулевую гипотезу.

В столбце $Var1$ показаны уровни фактора (1,2,3). Следующий столбец $Var2$ показывает групповую среднюю для каждого уровня фактора x_j . А столбец N – показывает количество испытаний на каждом уровне фактора.

Замечание 4.3. В только что рассмотренном примере на каждом уровне фактора было одинаковое число измерений. Но это число может быть и разным. И это ни в коей мере не усложняет процедуру дисперсионного анализа. Таков следующий пример.

Контрольный пример 4.2. В таблице приведены данные об урожайности сельскохозяйственной структуры за 6 лет при различных технологиях обработки почвы:

№ технологии	Годы					
	1-й	2-й	3-й	4-й	5-й	6-й
1	140	141	140	141	142	145
2	150	149	150	147		
3	147	147	145	150	150	
4	144	147	142	146		

Выяснить на уровне значимости $\alpha = 0{,}05$, зависит ли урожайность сельскохозяйственной культуры от технологии обработки почвы. Установить степень влияния технологии обработки почвы на урожайность.

Решение. Число уровней фактора $j = 4$; число наблюдений в группах $n_1 = 6$; $n_2 = 4$; $n_3 = 5$; $n_4 = 4$. Общее число наблюдений $n = 19$.

На рис. 4.12 приведены результаты проверки гипотезы о том, что урожайность не зависит от технологии обработки почвы ($\mathrm{M}\ X_1 = \cdots = \mathrm{M}\ X_4$) с помощью стандартных функций MS Excel.

№ технологии		F1	F2	F3	F4
Годы	1-й	140	150	147	144
	2-й	141	149	147	147
	3-й	140	150	145	142
	4-й	141	147	150	146
	5-й	142		150	
	6-й	145			
Групповые средние		141,5	149	147,8	144,75
Общая средняя		145,421	Общее число наблюдений		19
Групповые суммы квадратов отклонений					
		17,5	6	18,8	14,75
Общая сумма квадратов отклонений					
		230,632			

	Сумма квадратов отклонений	Число степеней свободы	Дисперсия	Fнабл	p-значение	Fкрит
межгрупповая	173,5816	3	57,8605	15,2131	0,0001	3,2874
внутригрупповая	57,05	15	3,8033			
общая	230,6316					
выборочный коэффициент детерминации			0,75			

Рис. 4.12 – Вид листа MS Excel с исходными данными и расчетами для дисперсионного анализа

Алгоритм решения данной задачи аналогичен описанному в контрольном примере 4.1.

Сделаем выводы. Так как $F_{\text{набл}} > F_{kp}$, то нулевая гипотеза отвергается, то есть на уровне значимости 0,05 (с надежностью 0,95 или 95%) выбор технологии существенно влияет на урожайность. Анализируя значение коэффициента детерминации, можем утверждать, что 75 % общего процента урожайности обусловлены технологией и лишь 25 % другими случайными

составляющими.

На рис. 4.13 представлено решение данной задачи с помощью инструмента *Однофакторный дисперсионный анализ* из *Пакета анализа*.

Однофакторный дисперсионный анализ

ИТОГИ

Группы	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
F1	6	849	141,5	3,5
F2	4	596	149	2
F3	5	739	147,8	4,7
F4	4	579	144,75	4,92

Дисперсионный анализ

Источник вариации	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
Между группами	173,582	3	57,861	15,213	8,07756E-05	3,287
Внутри групп	57,05	15	3,803			
Итого	230,632	18				

Рис. 4.13 – Результаты работы инструмента Однофакторный дисперсионный анализ

Решение в пакете STATISTICA.

Введём исходные данные из таблицы в созданную таблицу в формате STATISTICA, как показано на рисунке 4.14.

Data: Spreadsheet...

	1 Var1	2 Var2
1	1	140
2	1	141
3	1	140
4	1	141
5	1	142
6	1	145
7	2	150
8	2	149
9	2	150
10	2	147
11	3	147
12	3	147
13	3	145
14	3	150
15	3	150
16	4	144
17	4	147
18	4	142
19	4	146

Рис. 4.14 – Исходные данные. Var1 – факторы; Var2 – независимая переменная

Проведём анализ в модуле *ANOVA* (Дисперсионный анализ): в меню *Statistics* выберем *ANOVA*. На экране появится стартовая панель модуля. Выполним установки, как показано на рис. 4.15

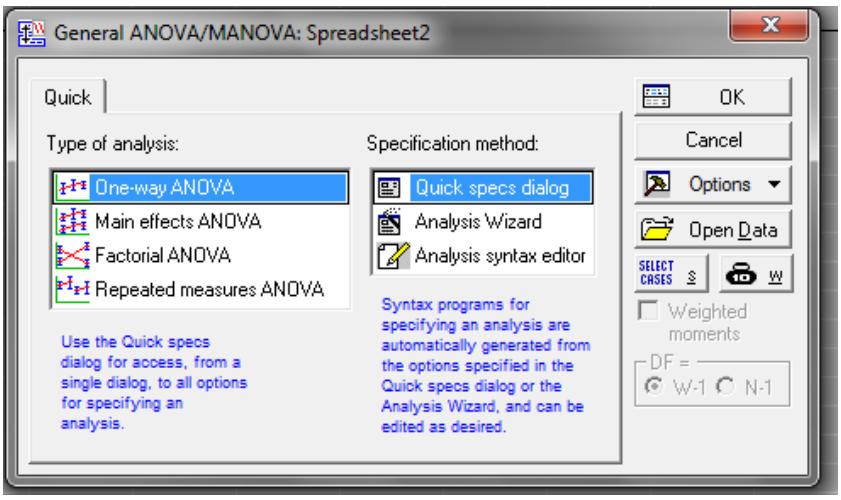


Рис. 4.15 – Стартовая панель модуля

После нажатия кнопки *OK* с помощью кнопки *Variables* (Переменные) в появившемся окне выберем переменные для анализа.

После того как кнопка будет нажата, на экране появится диалоговое окно *Select dependent variables and categorical predictor (factor)* (Выбрать списки зависимых переменных и факторов) (рисунок 4.16).

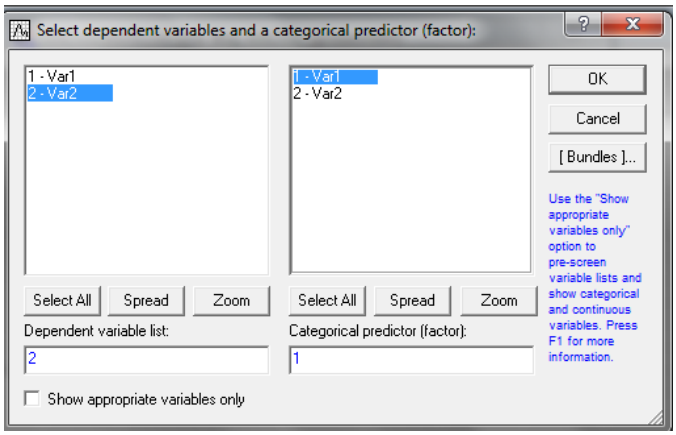


Рис. 4.16 – Окно выбора переменных для анализа

В левой части окна имя переменной выберем зависимую переменную, а в правой – фактор. После нажатия кнопки *OK* в появившемся окне выберем *All*, нажав кнопку *Factor codes* (рисунок 4.17).

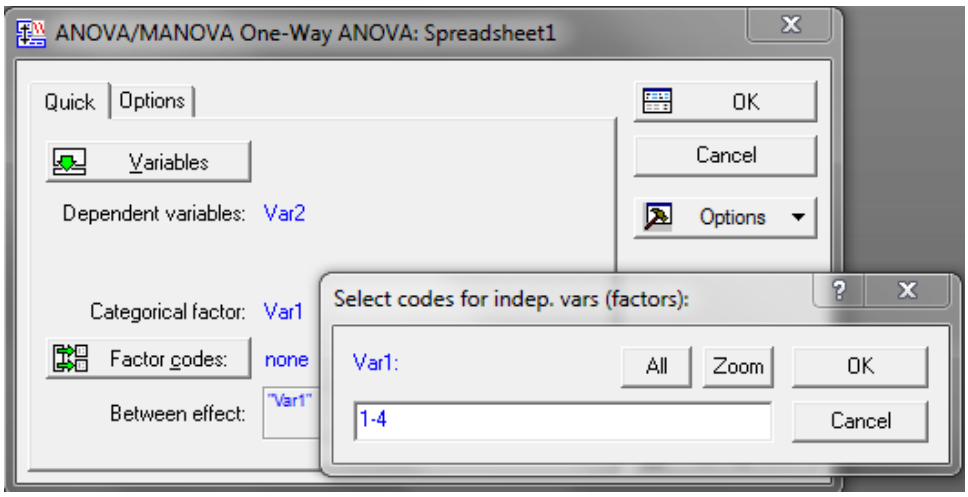


Рис. 4.17 – Окно выбора факторов

Нажмем кнопку *OK* в правом углу стартовой панели. На экране появится диалоговое окно *Anova Results* (Результаты) (рисунок 4.18). В данном окне на вкладке *Summary* выберем *Univariate Results* (Результат дисперсионного анализа). Далее нажмем кнопку *OK*.

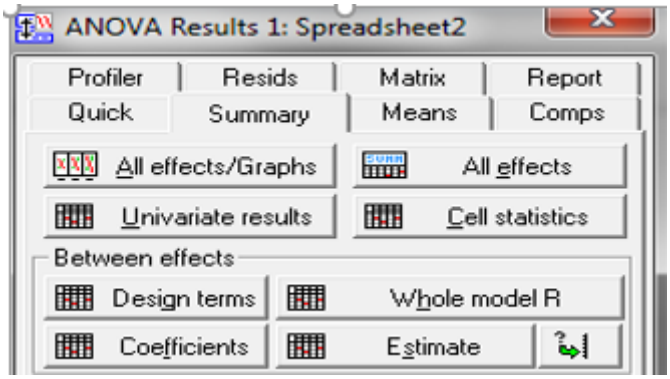


Рис.4.18 – Диалоговое окно результатов

В окне результатов (рис. 4.19) представлены результаты дисперсионного анализа:

между группами – *Var1*;

внутри групп – *Error*.

Univariate Results for Each DV (Spreadsheet1)
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	Degr. of Freedom	Var2 SS	Var2 MS	Var2 F	Var2 p
Intercept	1	392246,9	392246,9	103132,4	0,000000
"Var1"	3	173,6	57,9	15,2	0,000081
Error	15	57,0	3,8		
Total	18	230,6			

Рис. 4.19 – Результаты дисперсионного анализа

В рассмотренном примере *F-критерий* показывает, что различие между

средними статистически значимо (значимо на уровне 0,000081, 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*.

Контрольный пример 4.3. 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 $\overline{X}_i$
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 $\overline{X}_j$	**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:

	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. 4.20 – Исходные данные задачи

Выведем на экран диалоговое окно *Двухфакторный дисперсионный анализ без повторений*.

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.

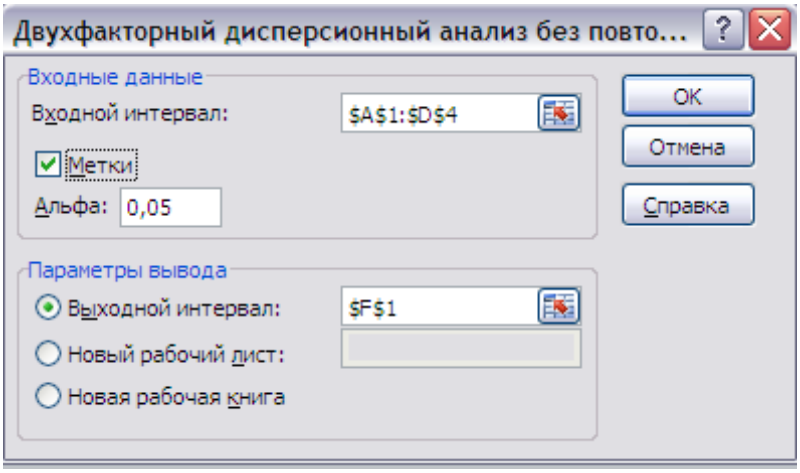


Рис. 4.21 – Диалоговое окно процедуры *Двухфакторный дисперсионный анализ без повторений.*

Щелкнем на кнопке ОК. Справа от таблицы исходных данных появятся 2 таблицы результатов рассматриваемой процедуры:

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. 4.22 – Результаты решения контрольного примера 4.3.

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:

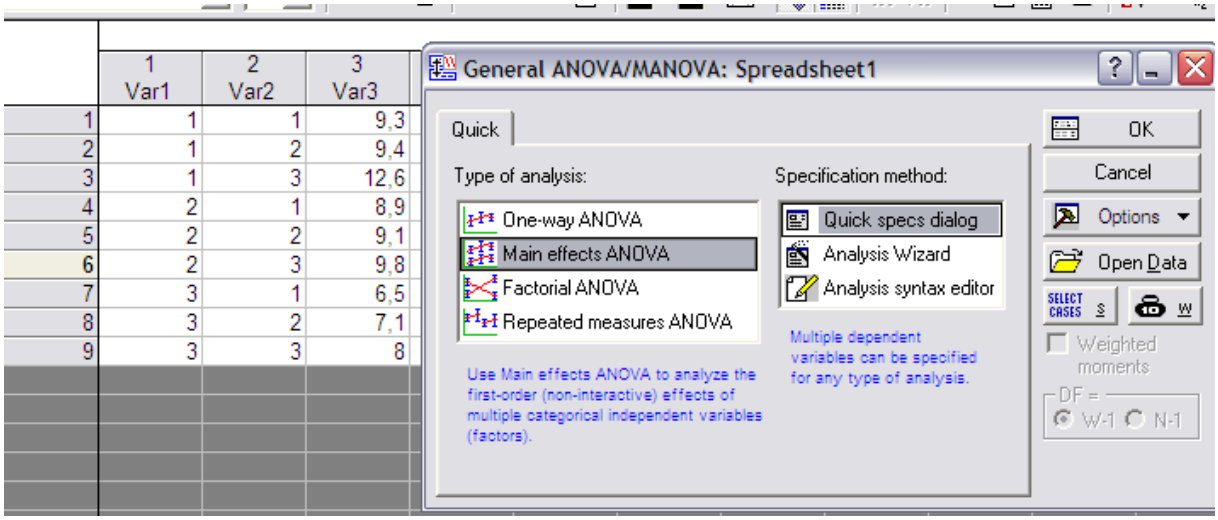


Рис. 4.23 – Исходные данные и вызов модуля.

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:

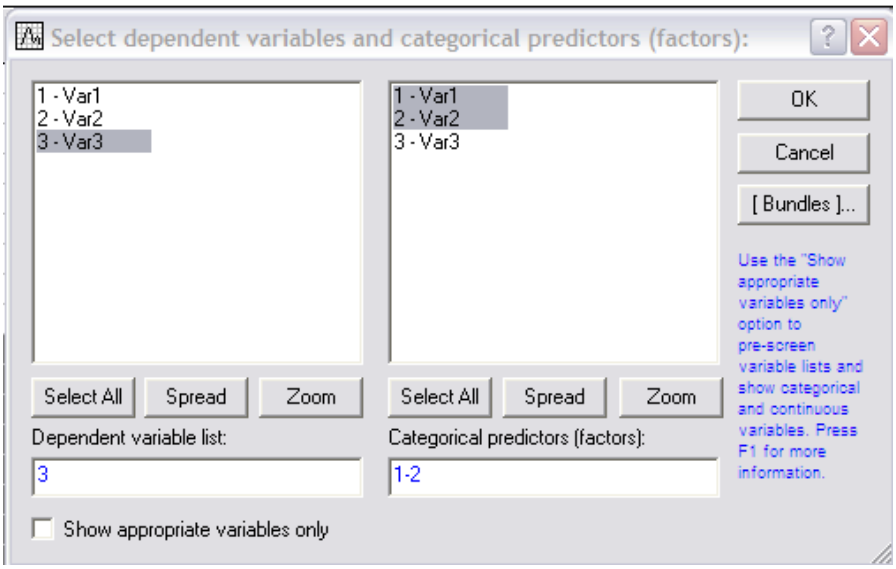


Рис. 4.24 – Выбор переменных.

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*:

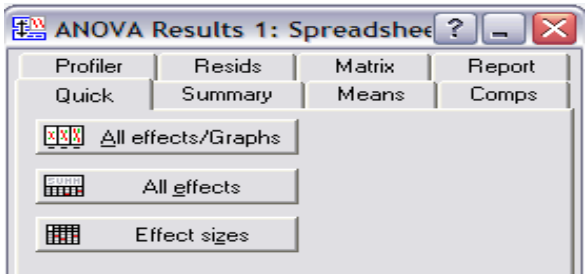


Рис. 4.25 – Панель *ANOVA Results*

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. 4.26 – 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

Контрольный пример 4.4. На 12 опытных делянках проводились экспериментальные работы с посевом кормовых злаковых трав. Факторы a и b отражают объективную ситуацию в процессе проведения опыта (a – освещённость, b – увлажнение) или фактор среды (неорганизованный фактор). Факторы o и p – факторы влияния: o – фоновые, без внесения удобрений, p – с внесением (организованный фактор). Результативным признаком является урожайность.

Исходные данные представлены в следующей таблице:

	a	***b***
o	58	49
o	84	55
o	39	48
p	72	74
p	72	74

p	64	85

Необходимо в пакетах *Excel* и *Statistica* провести двухфакторный дисперсионный анализ. Сделать вывод.

Решение. В проведении анализа в пакете *Excel* количество повторений (факторы *o* и *p*) должно быть одинаковым. Для выполнения подобного анализа в пакете *Statistica* количество повторений может быть различным.

Введём исходные данные (рис. 4.19). Откроем модуль *Анализ данных*, выберем процедуру *Двухфакторный дисперсионный анализ с повторениями*, после чего щёлкнем мышкой ОК. На экране появится диалоговое окно данной процедуры. Выполним операции и установки, как указано на рис. 4.27. Щёлкнем мышкой ОК.

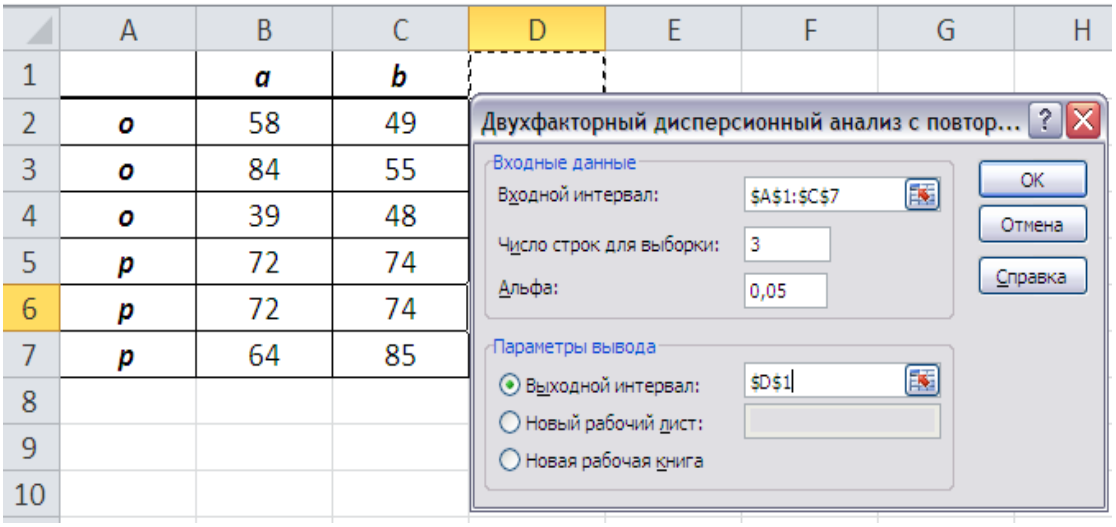


Рис. 4.27 – Исходные данные и диалоговое окно процедуры *Двухфакторный дисперсионный анализ с повторениями*

Результат обработки появится в указанном поле (рис. 4.28, 4.29):

23	Дисперсионный анализ						
24	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
25	Выборка	972	1	972	6,63104	0,032866806	5,317655072
26	Столбцы	1,333333	1	1,333333	0,009096	0,926363998	5,317655072
27	Взаимодействие	243	1	243	1,65776	0,233902095	5,317655072
28	Внутри	1172,667	8	146,5833			
29							
30	Итого	2389	11				

Рис. 4.28 – Дисперсионный анализ.

ИТОГИ	a	b	Итого
o			
Счет	3	3	6
Сумма	181	152	333
Среднее	60,33333	50,66667	55,5
Дисперсия	510,3333	14,33333	237,9
p			
Счет	3	3	6
Сумма	208	233	441
Среднее	69,33333	77,66667	73,5
Дисперсия	21,33333	40,33333	45,5
Итого			
Счет	6	6	
Сумма	389	385	
Среднее	64,83333	64,16667	
Дисперсия	236,9667	240,5667	

Рис. 4.29 – Статистические параметры

В рассмотренном примере критерий Фишера показывает, что нулевая гипотеза отвергается и различие между средними статистически значимо за счёт влияния второго фактора (строка *Выборка*) – значимо на уровне 0,033, что не превышает критического уровня 0,05. Сила влияния этого фактора определяется с помощью выборочного коэффициента детерминации и равна $\frac{972}{2389} = 0{,}4086$, т.е. составляет около 41%. В свою очередь, по первому фактору и взаимодействию обоих факторов нулевая гипотеза о равенстве средних не отвергается, поскольку критерий Фишера меньше табличного значения и $p > 0{,}05$. Поэтому, в данном случае прибавка к урожаю обусловлена только организованным фактором.

Введём исходные данные в созданную таблицу в формате *STATISTICA*, как показано на рисунке 4.30:

	1 Var1	2 Var2	3 Var3
1	a	o	58
2	a	o	84
3	a	o	39
4	a	p	72
5	a	p	72
6	a	p	64
7	b	o	49
8	b	o	55
9	b	o	48
10	b	p	74
11	b	p	74
12	b	p	85

Рис. 4.30 – Исходные данные
$Var3$ – независимая переменная
$Var1, Var2$ – факторы

Из переключателя модулей Statistica откроем модуль **ANOVA** (Дисперсионный анализ). На экране появится стартовая панель модуля (рис. 4.31).

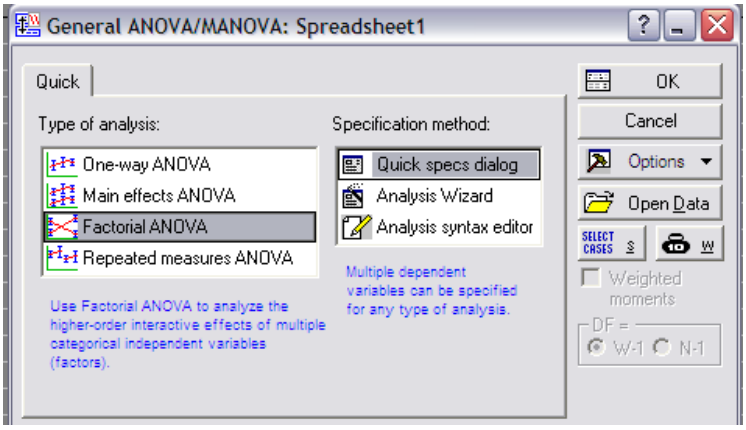


Рис. 4.31 – Стартовая панель модуля.

После нажатия кнопки OK в появившемся окне выберем переменные для анализа (рис. 4.32). Выбор переменных осуществляется с помощью кнопки *Variables* (Переменные).

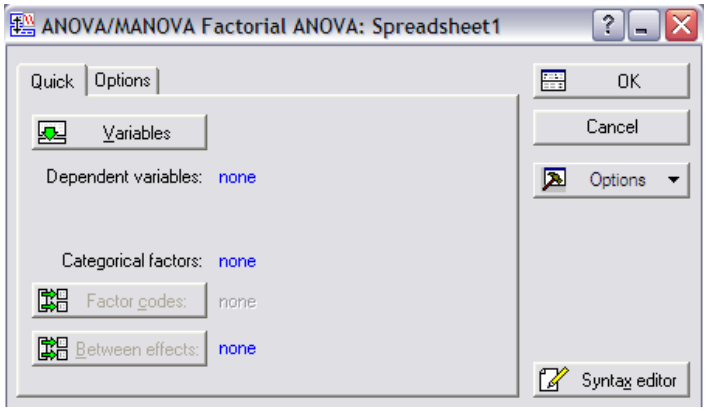


Рис. 4.32 – Окно выбора переменных

На экране появится диалоговое окно *Select depended variables and categorical predictor (factor)* (Выбрать списки зависимых переменных и факторов). В левой части окна выберем зависимую переменную, а в правой – факторы. После нажатия кнопки *OK* в появившемся окне выберем *All*, нажав кнопку *Factor code* (рис. 4.33).

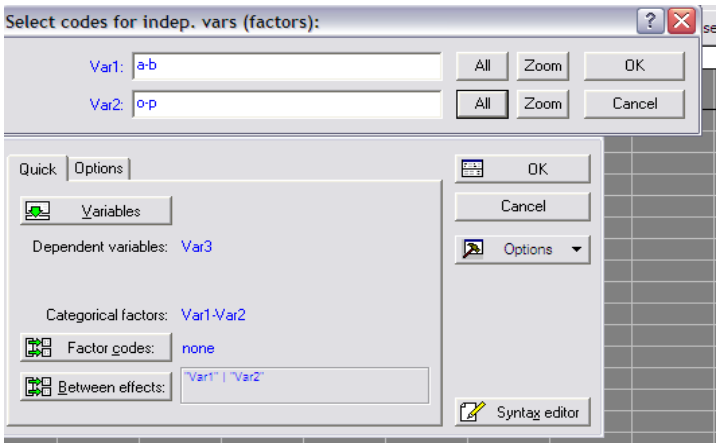
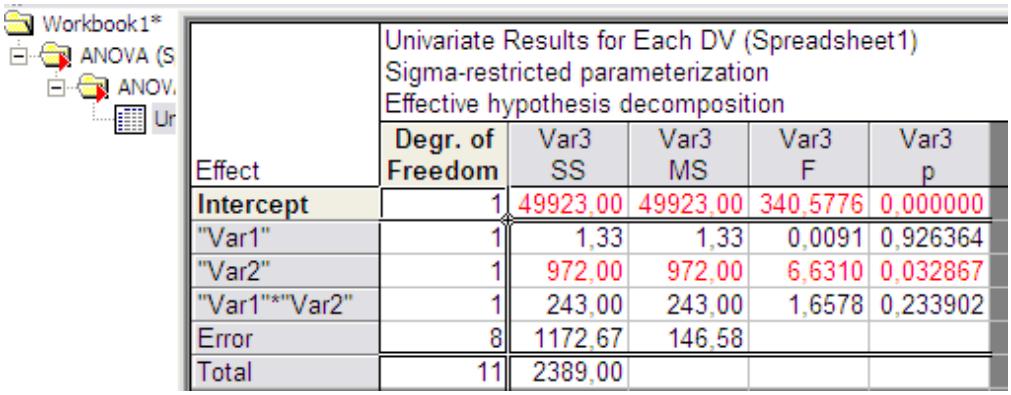


Рис. 4.33 – Окно выбора факторов

На экране появится диалоговое окно *Anova Results* (Результаты). В данном окне на вкладке Summary выберем *Univariate Result* (Результат дисперсионного анализа.

В окне (рис. 4.34) представлены результаты дисперсионного анализа:

- между группами, фактор 1 – *Var1*;
- между группами, фактор 2 – *Var2*;
- взаимодействие – *Var1* Var2*;
- внутри групп – *Error*.



Univariate Results for Each DV (Spreadsheet1)
Sigma-restricted parameterization
Effective hypothesis decomposition

Effect	Degr. of Freedom	Var3 SS	Var3 MS	Var3 F	Var3 p
Intercept	1	49923,00	49923,00	340,5776	0,000000
"Var1"	1	1,33	1,33	0,0091	0,926364
"Var2"	1	972,00	972,00	6,6310	0,032867
"Var1"*"Var2"	1	243,00	243,00	1,6578	0,233902
Error	8	1172,67	146,58		
Total	11	2389,00			

Рис. 4.34 – Результаты дисперсионного анализа

В рассмотренном примере критерий Фишера показывает, что различие между средними статистически значимо за счёт влияния второго фактора (на уровне 0,033) – результат в строке: между группами $Var2$ (фактор 2) подсвечивается красным цветом.

4.3. Задания для самостоятельной работы.

Задание 1.

Вариант 1. Произведено по 4 испытания на каждом из 4 уровней фактора F. Методом дисперсионного анализа при уровне значимости $\alpha = 0.05$ проверить нулевую гипотезу о равенстве известных групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с неизвестными дисперсиями.

№ испытания	Уровни фактора		
	F_1	F_2	F_3
1	31	24	21
2	30	26	22
3	35	20	34
4	32	30	31

Вариант 2. Выходной параметр — срок службы миниатюрного индикаторного прибора, ч. Уровни единственного фактора F — партии приборов, изготовленные по четырем разным технологиям. Отбор приборов для испытания полностью случаен. Было отобрано по 5 приборов из каждой партии. Проверить нулевую гипотезу о том, что варианты технологического процесса не влияют на срок службы индикаторных приборов. Принять $\alpha = 0{,}05$.

№ варианта технологического процесса	Приборы				
	1	2	3	4	5
F_1	1600	1610	1650	1680	1700
F_2	1580	1640	1640	1700	1750
F_3	1460	1550	1600	1620	1640
F_4	1510	1520	1530	1570	1600

Вариант 3. На предприятии, работающем в три смены, получены следующие данные о проценте брака выпускаемой продукции в каждой из смен за семь дней недели:

Смена	Дни недели						
	1	2	3	4	5	6	7
1	2	1,5	3	6	0,2	0	1
2	1,5	4	4	0	0	2,5	2,5
3	1,5	1,5	6	6	0	3	1

Методом дисперсионного анализа проверить нулевую гипотезу: эффект влияния фактора F – смены на процент брака отсутствует. Принять $\alpha = 0{,}05$.

Вариант 4. Исследовалось влияние четырех различных типов покрытия на удельную проводимость телевизионных трубок. Результаты наблюдений:

Наблюдения	Типы покрытия			
	1	2	3	4
1	56	64	45	42
2	55	61	46	39
3	52	50	45	45
4	59	55	39	43
5	60	56	43	41

Принять $\alpha = 0{,}05$.

Вариант 5. На трех станках изготавливаются детали одного и того же размера. Из продукции каждого было отобрано по четыре образца. Предполагается, что выборки сделаны из нормальных совокупностей с одинаковыми дисперсиями.

Можно ли утверждать, что партии изготовляемых деталей имеют одинаковые групповые средние. Гипотезу проверить при $\alpha = 0{,}05$.

Номер станка	Номера деталей			
	1	2	3	4
1	38	36	35	31
2	20	24	26	30
3	21	22	31	34

Вариант 6. Требуется при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей с одинаковыми генеральными дисперсиями.

№ испытания	Уровни фактора			
	F_1	F_2	F_3	F_4
1	6	6	9	7
2	7	7	12	9
3	8	11	13	10
4	11	12	14	10

Вариант 7. Для проверки влияния внутрицехового оформления на каче-

ство продукции рассмотрены три участка по производству однотипной продукции и произведена выборочная проверка процента брака за 5 месяцев. Методом дисперсионного анализа при уровне значимости $\alpha = 0{,}05$ проверить нулевую гипотезу о существенности влияния оформления участка (фактор F) на качество продукции. Результаты проверки приведены в таблице:

Номер измерения	Уровни фактора		
	F_1	F_2	F_3
1	3	5	1
2	2	4	4
3	1	3	5
4	2	6	10
5	4	3	3

Вариант 8. 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.

Тип рекламы	Годы					
	1	2	3	4	5	6
A	10759	11248	11778	10557	11109	10860
B	11243	11283	10617	10889	10406	10363
C	11904	11315	11852	11268	11394	11080
D	11336	12584	12582	12241	11900	12390

Задание 2.

Вариант 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.

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 2. Требуется при уровне значимости 0,05 проверить нулевую

гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей.

Номер испытания	Уровни фактора		
	F_1	F_2	F_3
1	3.7	6	6.9
2	4.7	8.6	10
3	4	6.7	9.8
4	6	9.2	
5		9.5	
6		9.8	

Вычислить коэффициент детерминации, сделать вывод

Вариант 3. Требуется при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей.

Номер испытания	Уровни фактора		
	F_1	F_2	F_3
1	30.56	43.44	31.36
2	32.66	47.51	36.2
3	34.78	53.8	36.38
4	35.5		42.2
5	36.63		
6	40.20		
7	42.28		

Вычислить коэффициент детерминации, сделать вывод.

Вариант 4. Исследователь хочет узнать, если ли разница в прибавке веса

у спортсменов, следующих специальной диете. Спортсмены разделены на три группы случайным образом, каждая группа следует определенной диете 6 недель. Прибавка в весе указана в таблице.

Диета № 1	Диета № 2	Диета № 3
3	10	8
4	12	3
7	11	2
4	14	5
	8	
	6	

На уровне значимости $\alpha = 0{,}05$ определить, может ли исследователь утверждать, что есть разница в диетах?

Вычислить коэффициент детерминации, сделать вывод.

Вариант 5. Три мастера проводили проверку однотипных устройств.

Каждый из них проверил различное число комплектов и обнаружил различное число дефектов при каждой проверке. Анализируя их работу, начальник наладочного участка составил таблицу, из которой можно предположить, что 2-й мастер, вроде бы, имеет более низкую квалификацию или у него отсутствует служебное рвение, т.к. он обнаружил в среднем в два раза меньше дефектов (от общего среднего). Можно ли на основании таблицы сделать вывод о различной квалификации мастеров?

Проверка	Мастер		
	1	2	3
1	11	6	8
2	7	1	7
3	8	2	9
4	4		4
5	5		

Вычислить коэффициент детерминации, сделать вывод.

Вариант 6. Испытывались на долговечность электрические лампы. Вы-

борки были взяты из 4-х партий, изготовленных из разных материалов.

Партии	Продолжительность горения в часах							
	1	2	3	4	5	6	7	8
1	1600	1610	1650	1680	1700	1700	1800	
2	1580	1640	1700	1750				
3	1460	1550	1600	1620	1640	1660	1740	1820
4	1510	1520	1530	1570	1600	1680		

Методом дисперсионного анализа проверить нулевую гипотезу о том, что средние долговечности ламп в партиях одинаковы и не зависят от применен-ных материалов. Вычислить коэффициент детерминации, сделать вывод

Вариант 7. Требуется при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей.

Номер испытания	Уровни фактора				
	F_1	F_2	F_3	F_4	F_5
1	7.3	5.4	6.4	7.9	7.1
2	7.6	7.1	8.1	9.5	
3	8.3	7.4		9.6	
4	8.3				
5	8.4				

Вычислить коэффициент детерминации, сделать вывод.

Вариант 7. Требуется при уровне значимости 0,05 проверить нулевую гипотезу о равенстве групповых средних. Предполагается, что выборки извлечены из нормальных совокупностей. Вычислить коэффициент детерминации, сделать вывод

Уровень фактора	Номер измерения					
	1	2	3	4	5	6
F_1	21,6	20,5	23,4	22,1	21,7	21,6
F_2	21,9	23,0	22,5	21,3	22,7	22,8
F_3	22,8	22,5	23,4	21,8	22,2	
F_4	24,1	24,4	21,9	24,1		

Задание 3.

Вариант 1. Урожай зерновых культур X (ц/га) в зависимости от срока посева A (оптимальный и поздний) и от сорта B дан в таблице (в строке AB – первый индекс – уровень фактора A, второй – уровень фактора B):

№ п/п	1	2	3	4
AB	1 1	1 2	2 1	2 2
x_{ijr}	41; 42; 37; 44	35; 38; 37; 34	36; 32; 37; 35	30; 31; 34; 33

Проверить влияние факторов A и B на урожайность X. Принять $\alpha = 0{,}05$.

Вариант 2. При исследовании зависимости товарооборота центральной районной аптеки от товарооборота A и штатной численности прикреплённой аптечной сети B получен двухфакторный комплекс. При $\alpha = 0{,}05$ проверить влияние факторов A и B на товарооборот.

Фактор B	Фактор A		
	A_1	A_2	A_3
B_1	157	163	161
B_2	160	165	158
B_3	158	163	158

Вариант 3. В таблице представлены данные об урожайности (ц/га) четырёх сортов пшеницы (четыре уровня фактора A), достигнутой при использовании пяти типов удобрений (пять уровней фактора B).

Фактор B – тип удобрения	Фактор A – сорт пшеницы			
	A_1	A_2	A_3	A_4
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

Данные получены на 20 участках одинакового размера и аналогичного почвенного покрова. Необходимо определить, влияет ли сорт и тип удобрения на урожайность пшеницы. Принять $\alpha = 0{,}05$.

Вариант 4. В двухфакторном комплексе проводится 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 рабочих.

Фактор B	Фактор A		
	A_1	A_2	A_3
B_1	195	198	202
B_2	196	201	203
B_3	198	202	204

Вариант 5. Пусть экспериментально проверяется влияние на износоустойчивость детали таких факторов, как материал (два вида) и технология изготовления (три метода). Данные экспериментов (число месяцев работы детали) приведены в таблице:

Материал (фактор B)	Технология (фактор A)		
	1	2	3
1	10; 8; 7; 10	8; 12; 14; 12	18; 8; 10; 10
2	12; 8; 8; 7	12; 13; 11; 14	13; 15; 12; 10

Предполагается, что уровни факторов A и B фиксированные, а число месяцев работы детали есть случайная величина, распределённая по нормальному закону распределения. Проверить при $\alpha = 0{,}05$ существенность влияния на число месяцев работы детали: а) материала; б) методики изготовления; в) взаимодействия факторов AB .

Вариант 6. Медработник хочет проверить влияние двух разных диет и двух типов упражнений на уровень глюкозы в крови. Уровень глюкозы измеряется в миллиграммах на децилитр (mg/dl). Для этого в каждую из групп были распределены по три человека. Проанализируйте нижеприведённые данные, используя двухфакторный дисперсионный анализ при $\alpha = 0{,}05$.

Упражнения	Диета A	Диета B
I	62; 64; 66	58; 62; 63
II	65; 68; 72	83; 85; 91

Вариант 7. Компания-производитель желает проверить эффективность различных видов рекламы. Для рекламируемого продукта созданы два типа рекламных роликов: серьёзный и смешной. Ролики размещены в рабочие и выходные дни:

Тип ролика	Рабочий день	Выходной день
Смешной	6; 10; 11; 9	15; 18; 14; 16
Серьёзный	8; 13; 12; 10	19; 20; 13; 17

Выбраны 16 потенциальных покупателей и случайным образом распределены на 4 группы. После того как каждый покупатель просмотрел ролик, его просят оценить рекламу по шкале из 20 баллов. Различные баллы даются за привлекательность, ясность, краткость ролика и т.д. При $\alpha = 0{,}01$ требуется проанализировать данные (используя двухфакторный дисперсионный анализ) и сделать выводы..

Вариант 8. 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:

Уровни фактора A	Уровни фактора B				
	B_1	B_2	B_3	B_4	B_5
A_1	38	32	46	44	35
A_2	44	41	45	42	33
A_3	32	33	40	37	33
A_4	31	36	38	36	34

Методом дисперсионного анализа проверить гипотезу о том, что факторы A и B не влияют на математическое ожидание случайной величины X. Принять $\alpha = 0{,}05$.

Лабораторная работа № 5

КОРРЕЛЯЦИОННЫЙ АНАЛИЗ

Цель работы: Овладение методами исследования корреляционной зависимости между несколькими количественными случайными величинами по выборочным данным.

Используемые программные средства: MS Excel 2010 (2016), STATISTICA 8.0.

5.1. Краткие теоретические сведения.

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 (см. рис. 5.1), т.е. оценить *направленность* статистической зависимости.

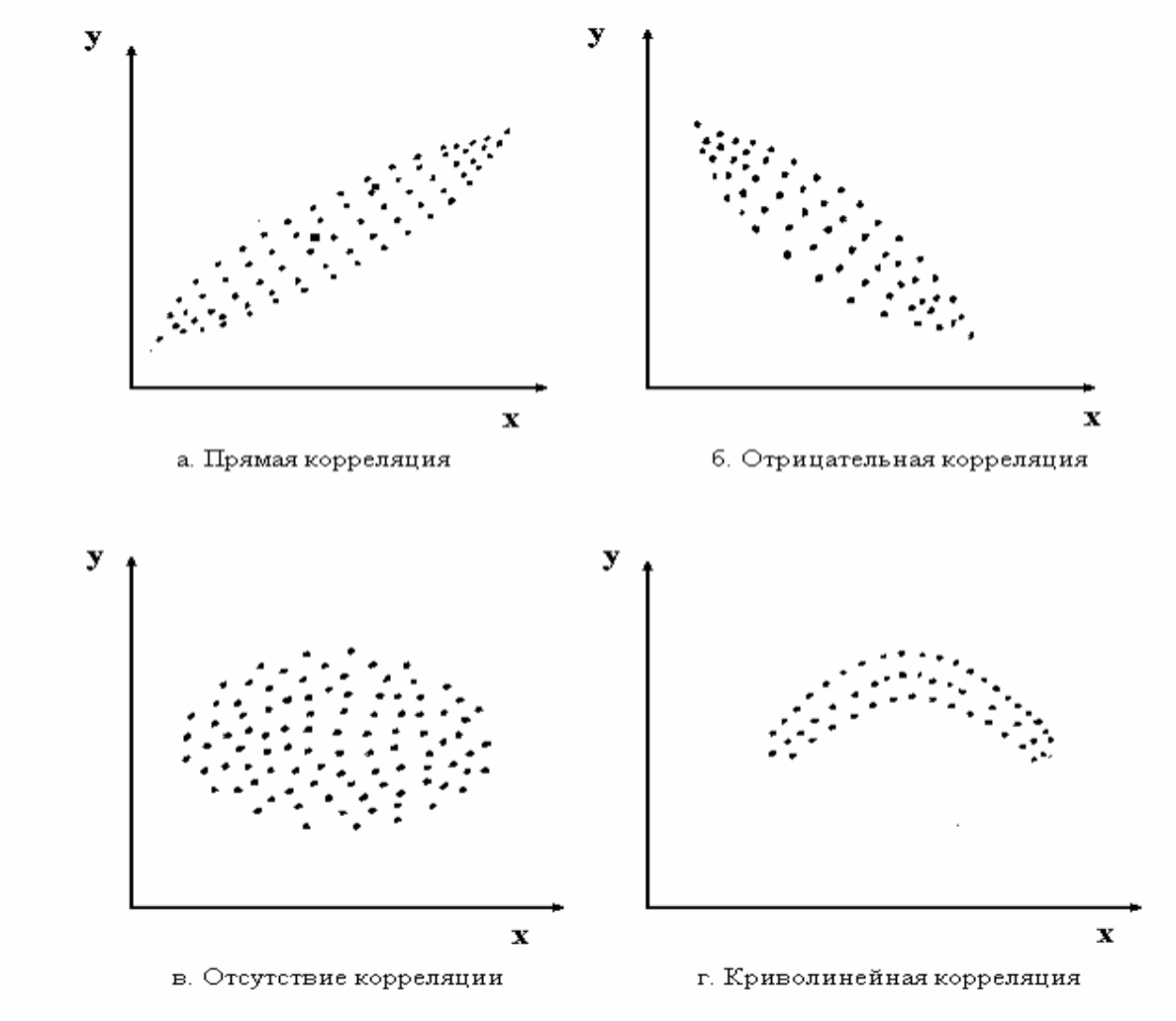


Рис. 5.1 – Виды корреляционных полей

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

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

$$K_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} x_i - x \quad y_i - y \tag{5.1}$$

где $x = \frac{1}{n} \sum_{i=1}^{n} x_i$; $y = \frac{1}{n} \sum_{i=1}^{n} y_i$ – 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 величи-

на) – *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:

$$r_B = \frac{cK_{XY}}{s_x s_y} = \frac{\sum_{i=1}^{n} x_i - x \quad y_i - y}{\sum_{i=1}^{n} x_i - x \ ^2 \ y_i - y \ ^2} \tag{5.2}$$

Коэффициент линейной корреляции всегда удовлетворяет соотношению:

$$-1 \leq r_B \leq 1.$$

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 $-1 < r_B < 0$, 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. 5.1):

Таблица 5.1

r_{XY}	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 кор-

релирует 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

$$T = \frac{r_B \sqrt{n-2}}{\sqrt{1-r_B^2}} \tag{5.4}$$

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:

$$r_B - t\ \alpha, n-2\ \cdot \sigma_{r_B} < r_{XY} < r_B + t\ \alpha, n-2\ \cdot \sigma_{r_B} \tag{5.5}$$

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.

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

$$\rho_B = 1 - \frac{6 \cdot S}{n^3 - n}$$

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

$$T = \frac{\rho_B \sqrt{n-2}}{\sqrt{1-\rho_B^2}}$$

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 5.2:

Таблица 5.2

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:

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

Выборочный коэффициент ранговой корреляции Кендалла находится по

формуле:

$$\tau_B = \frac{4R}{n\ n-1} - 1 = \frac{2\ R-Q}{n\ n-1}$$

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).

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:

$\boldsymbol{X}$	$\boldsymbol{Y}$						
	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}	$\boldsymbol{n}$

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. 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:

$$\eta = \frac{\sigma_{\text{межгр}}}{\sigma_{\text{общ}}}$$

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

$\sigma^2_{\text{межгр}} = \frac{\sum_{i=1}^{l} n_{xi}\ y_{xi} - y\ ^2}{n}$ – межгрупповая дисперсия, характеризует разброс условных средних y_{xi} от общей средней y; $\sigma^2_{\text{общ}} = \sigma^2_y$ – общая дисперсия, характеризует разброс фактических данных y_j от их общей средней y.

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

$\sigma^2_{\text{межгр}} = \frac{\sum_{j=1}^{m} n_{yj}\ x_{yj} - x\ ^2}{n}$ – межгрупповая дисперсия, характеризует разброс условных средних x_{yj} от общей средней x, $\sigma^2_{\text{общ}} = \sigma^2_x$ – общая дисперсия, характеризует разброс фактических данных x_i от их общей средней x.

Корреляционное отношение обладает следующими свойствами.

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

$$r_B\ \le \eta_{y/x}; \quad r_B\ \le \eta_{x/y}.$$

5.2. Практическая часть.

Контрольный пример 5.1. Для исследования зависимости случайных величин X и Y получены статистические данные, представленные в таблице (табл. 5.3).

Таблица 5.3.

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:

a) Построить корреляционное поле, сделать вывод.
b) Найти выборочный коэффициент корреляции.
c) При уровне значимости $\alpha = 0.05$ проверить нулевую гипотезу о равенстве генерального коэффициента корреляции нулю при конкурирующей гипотезе $H_1: r_{XY} \neq 0$.

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. 5.2).

	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. 5.2 – Исходные данные

Построим корреляционное поле, выбрав в меню *Graphs* (Графики) пункт *Scatterplots* (Диаграммы рассеяния). Далее выбираем вкладку *Advanced:*

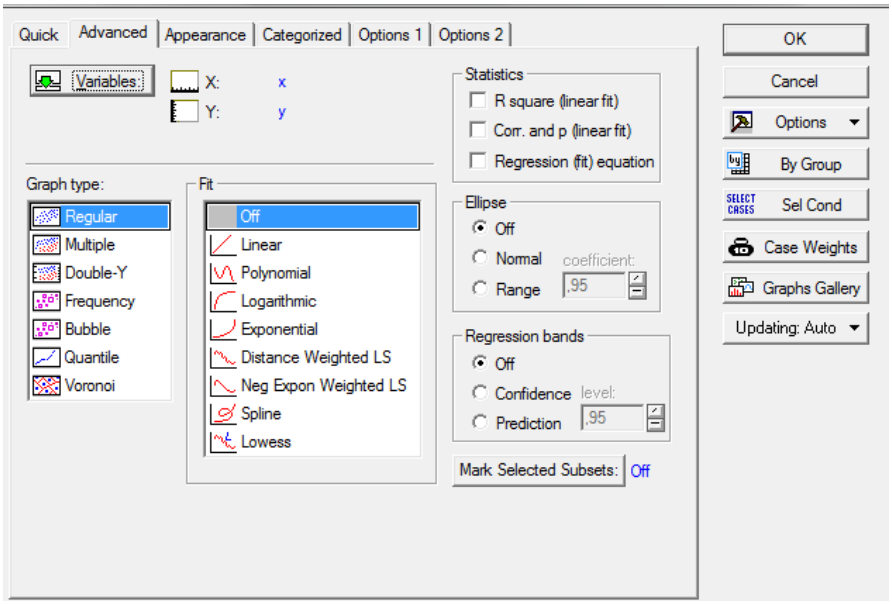


Рис. 5.3 – Вкладка *Advanced* меню *Graphs – Scatterplots*

Выберем *Graph type/Regular*; *Fit/Off* и нажмём кнопку *OK*, после чего на экране появляется график корреляционного поля (рис. 5.4):

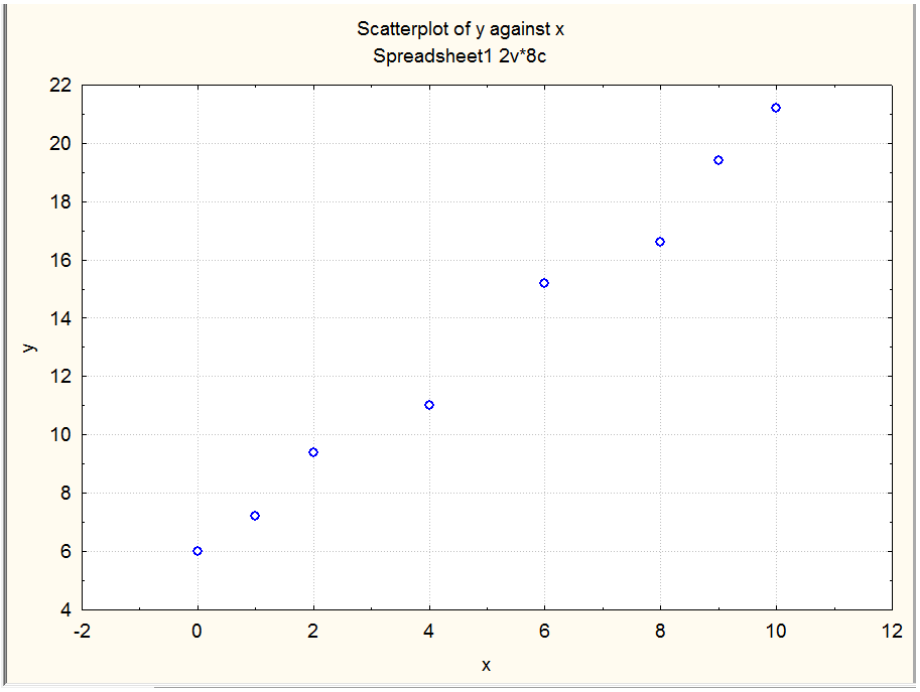


Рис.5.4 – Корреляционное поле

Далее открываем меню *Statistics* и выбираем пункт *Basic Statistics and Tables* (Основные статистики). После этого в появившемся окне открываем пункт *Correlation matrices* (Корреляционная матрица).

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.5.5).

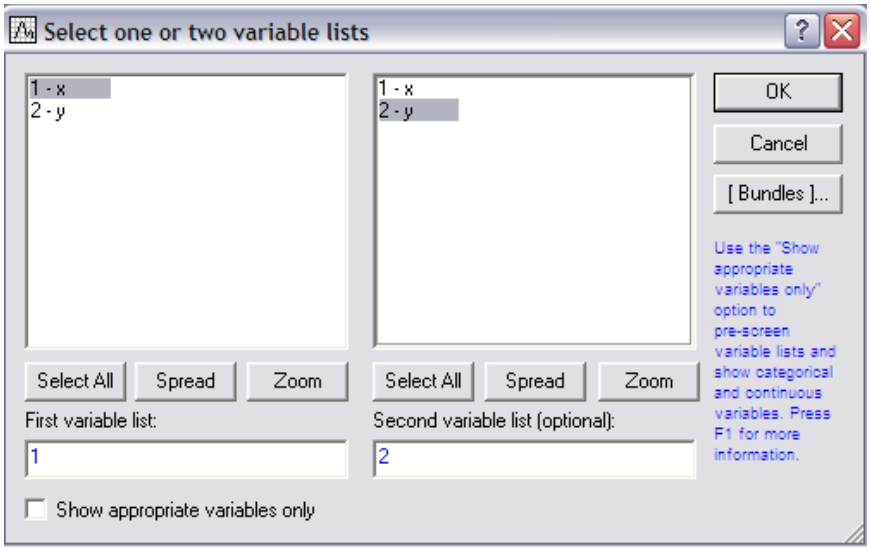


Рис. 5.5 – Выбор переменных для корреляционного анализа

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. 5.6):

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. 5.6 – 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. 5.7). Зададим команду *Данные – Анализ данных* и выберем инструмент *Корреляция*. Заполним диалоговое окно «Корреляция» как показано на рис. 5.7.

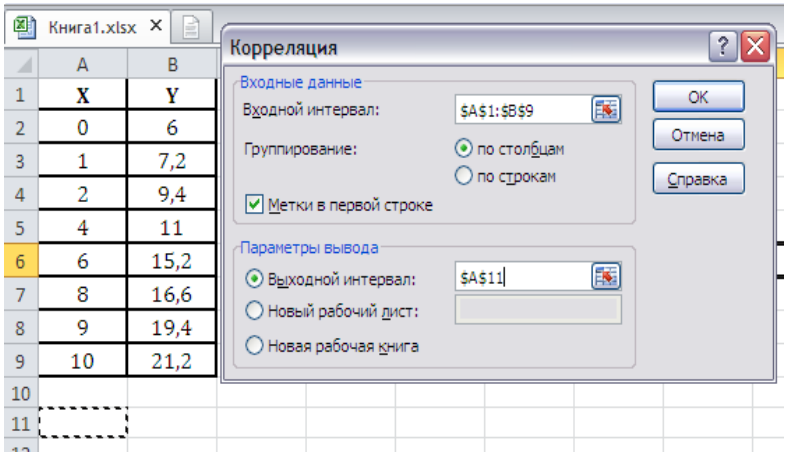


Рис. 5.7 – Исходные данные и пример заполнения диалогового окна «Корреляция»

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. 5.8), что свидетельствует о весьма высокой степени прямой линейной связи (по шкале Чаддока). Проверим его на значимость. Найдём наблюдаемое значение T – статистики по формуле (5.4).

Введём в ячейку E11 формулу MS Excel (рис. 5.8):

$$= \text{ABS(B13)/КОРЕНЬ}(1 - \text{B13\^{}2}) * \text{КОРЕНЬ}(8 - 2).$$

Для расчёта критического значения T – статистики при уровне значимости $\alpha = 0.05$ и числе степеней свободы $n - 2 = 6$ введём в ячейку E13 формулу $= \text{СТЬЮДЕНТ.ОБР.2X}(0{,}05; 6)$. Результаты показаны на рис. 5.8.

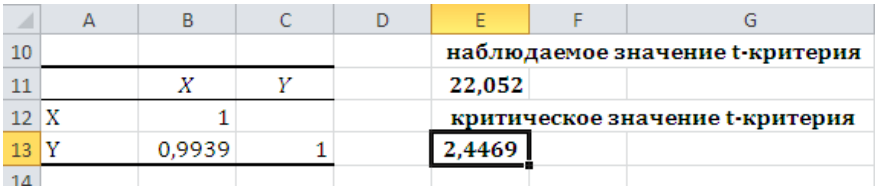


Рис. 5.8 – Корреляционная матрица, сформированная *Пакетом анализа* и наблюдаемые и критические значения T – статистики

Так как $T_{\text{набл}} > t_{кр}$ – отвергаем нулевую гипотезу о равенстве нулю генерального коэффициента корреляции. Значит, X и Y линейно коррелированы.

Замечание 5.1. В пакете *Excel* для вычисления выборочных коэффициента корреляции и корреляционного момента можно использовать стандартные

функции (см. рис. 5.9), 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. 5.9 – Расчет коэффициента корреляции и корреляционного момента с помощью стандартных функций пакета *Excel*.

Контрольный пример 5.2. 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).

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.5.10).

	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.5.10 – 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 5.11). 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.

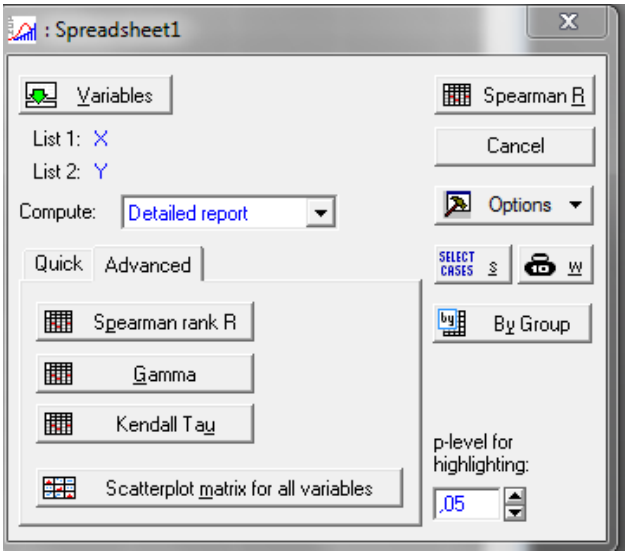


Рис.5.11 – Меню настроек для вычисления коэффициента Спирмена

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.5.12).

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.5.12 – 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. 5.13).

Pair of Variables	Kendall Tau Correlations (Spreadsheet1) MD pairwise deleted Marked correlations are significant at p <,05000 Valid N	Kendall Tau	Z	p-level	p-exact 1-tailed
X & Y	12	0,787879	3,565773	0,000363	----

Рис.5.13 – Коэффициент ранговой корреляции Кендалла

Здесь $p = 0{,}000363 < 0{,}05$.

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

	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. 5.14 – 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.

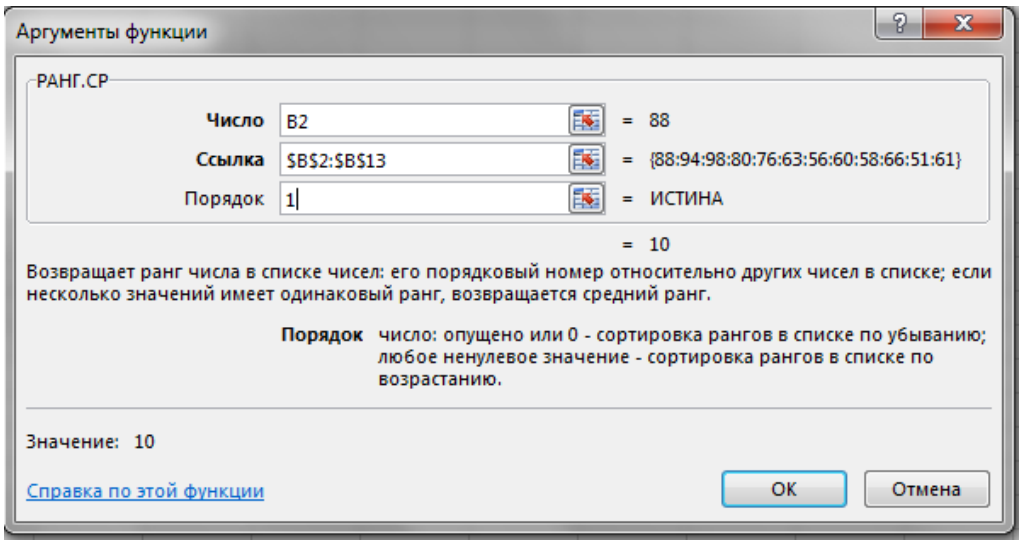


Рис. 5.15 – Диалоговое окно функции РАНГ.СР

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.

Выделим диапазон G2:G13 и перейдя на вкладку *Формулы* выберем инструмент ***Автосумма***.

	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. 5.16 – Расчёт коэффициента ранговой корреляции Спирмена

Замечание 5.2. 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 5.3. В ячейке G15 записана выборочная оценка коэффициента

ранговой корреляции Спирмена, вычисленная по формуле $= 1 - 6 * G14/(12\text{^}3 - 12)$.

Замечание 5.4. В ячейках J2: J4 находятся наблюдаемое значение t статистики T, её критическое значение (критическое значение распределения Стьюдента) и значимость p (см. рис. 5.17):

	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. 5.17 – Формулы для расчёта наблюдаемого значения статистики T, её критического значения и значимости p

Так как $t > t\ 0{,}025; 10$ и $p \ll \alpha$ – 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. 5.18, диапазон A1:E13).

	A	B	C	D	E	F	G	H
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. 5.18 – Исходные данные и решение контрольного примера 6.2 (коэффициент Кендалла)

2. Выделим диапазон, в котором находятся ранги оценок и щелкнем на кнопке «*Копировать*».

3. Выделим ячейку F1, выберем в меню *Главная* – *Вставить* – *Специаль-*

ная вставка – Значения. В диапазоне F2:G13 появятся «копии» рангов экспертных оценок.

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 $= \text{СУММ}(\text{ЕСЛИ}(\$G3{:}\$G\$13 > G2; 1; 0))$, нажмём клавиши $Ctrl - Shift - Enter$ 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 $= 4 * \text{H14}/12/11 - 1$, 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 $= \text{НОРМ.ОБР}(1 - 0{,}05/2; 0; 1)$ (рис. 5.19, ячейка K2)

2) Найдем критическую точку (рис. 5.19, ячейка K3) по формуле

$$T_{\text{кр}} = z_{\text{кр}} \sqrt{\frac{2\ (2n+5)}{9n\ (n-1)}} = 1{,}96 \cdot \sqrt{\frac{2 \cdot\ (2 \cdot 12 + 5)}{9 \cdot 12 \cdot 11}} = 0{,}433066$$

J	K	L	M	N	O	P
zkp	**1,959964**					
Tkp	**0,433066**					
H0 отвергают - ранговая корреляционная связь является значимой						

Рис. 5.19 – Проверка коэффициента ранговой корреляции Кендалла на значимость

Так как $\tau_{\text{B}} > \text{T}_{\text{кр}}$ – 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 5.3***.** В табл. 5.4 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 5.4

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. 5.20):

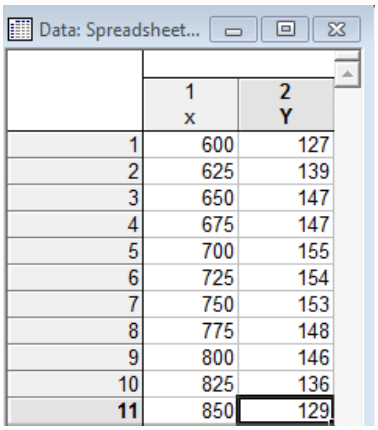


Рис. 5.20 – Исходные данные

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. 5.21):

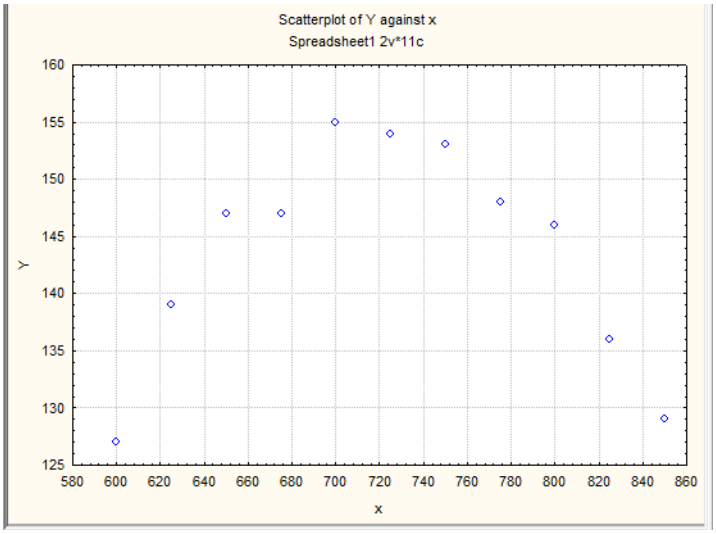


Рис. 5.21 – Диаграмма рассеивания

Корреляционное поле, показанное на рис. 5.21, 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, используя пункт *Correlation matrices* меню *Statistics* (см. контрольный пример 5.1):

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	N
x	725,0000	82,91562					
Y	143,7273	9,70661	-0,015531	0,000241	-0,046599	0,963850	11

Рис. 5.22 – 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 5.5: 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. 5.23. 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 гв**	**-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. 5.22 – Вид листа *MS Excel* с исходными данными и расчетами для вычисления корреляционного отношения

Для каждой группы рассчитаем *групповую среднюю*. В ячейку B11 введем формулу $=\text{СРЗНАЧ(B7:B9)}$, 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 $=\text{КОРРЕЛ(B1:L1;B2:L2)}$.

Рассчитаем квадраты отклонений групповых средних от общей средней (диапазон B12:E12). Введем в ячейку B12 формулу $=\text{(B11}-\text{\$E\$4)\^{}2}$, которую затем скопируем вправо по строке.

Межгрупповую дисперсию рассчитаем в ячейке K7 по формуле $=\text{СУММПРОИЗВ(B12:F12;B10:F10)/11}$, 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 $=\text{КОРЕНЬ(K7/K6)}$. Полученное значение свидетельствует о наличии сильного нелинейного влияния температуры на объем выпуска продукции.

5.3. Задания для самостоятельной работы.

Задание 1. Для исследования зависимости случайных величин X и Y получены статистические данные.

Требуется:

- Построить корреляционное поле, сделать вывод.
- Найти выборочный коэффициент корреляции.
- При уровне значимости $\alpha = 0.05$ проверить нулевую гипотезу о равенстве генерального коэффициента корреляции нулю при конкурирующей гипотезе $H_1: r_{XY} \neq 0$.

1	X	-1	-0,75	-0,5	-0,25	0	0,25	0,5	0,75	1	
	Y	2,08	1,83	1,57	1,13	0,89	0,75	0,3	0,06	-0,01	
2	X	0,95	1,21	1,47	1,74	2	2,26	2,52	2,78	3,05	3,31
	Y	3,16	2,39	2,19	1,34	1	0,8	0,54	0,4	0,28	0,19
3	X	-5	-3,91	-2,82	-1,73	-0,64	-0,45	1,54	2,63	3,72	
	Y	0	-0,01	-0,02	-0,03	-0,05	-0,1	-0,13	-0,2	-0,24	
4	X	-5	-4	-3	-2	-1	0	1	2	3	4
	Y	0,04	0,05	0,08	0,12	0,16	0,21	0,28	0,36	0,4	0,45
5	X	1,2	1,6	2,5	2,7	3,1	3,5	4,3	4,9	5,5	6,4
	Y	1,2	1,5	1,8	2,7	3,4	4,3	5,8	7,45	8,34	10,5
6	X	-1	-0,7	-0,4	-0,2	0,1	0,4	0,7	1	1,2	1,5
	Y	-5	-4,9	-4,7	-4,4	-3,6	-1,9	-1	0,1	0,3	0,6
7	X	0,01	0,53	1,05	1,57	2,09	2,61	3,12	3,64	4,16	4,68
	Y	5,2	3,3	2,26	1,05	-0,8	-1,7	-2,2	-2,5	-2,9	-3,2
8	X	11	12	13	14	15	16	17	18	19	
	Y	115	121	132	134	145	155	164	172	182	

Задание 2. Два преподавателя A и B 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 $\alpha = 0{,}05$.

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

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*).

1	X	-2,1	-1,8	-1,5	-1,2	-0,9	-0,6	-0,2	0,1	0,4	0,7
	Y	0,28	0,29	0,3	0,32	0,36	0,48	0,78	1,52	3,41	8,21
2	X	1	2	3	4	5	6	7	8	9	10
	Y	7,5	5,5	4	3	2	1,5	1	1	0,5	0,5
3	X	0,01	0,51	1,01	1,52	2,01	2,51	3	3,05	4	4,5
	Y	-1,14	2,39	3,01	3,37	3,63	3,83	3,99	4,13	4,25	4,35
4	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
5	X	-2	-1,5	-1	-0,5	0	0,5	1	1,5	2	2,5
	Y	19,9	11,1	5,4	2,8	3,3	6,5	13	23	36	52
6	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
7	X	-1	-0,5	0	0,5	1	1,5	2	2,5	3	3,5
	Y	1	0,5	0,3	0,6	1,3	2,6	4,3	6,6	9,3	12,6
8	X	0,5	0,6	0,7	0,8	0,9	1	1,1	1,2	1,3	1,4
	Y	0,705	0,495	0,426	0,357	0,368	0,406	0,549	0,768	1,012	1,331

Лабораторная работа № 6

РЕГРЕССИОННЫЙ АНАЛИЗ

Цель работы: привить навыки по анализу установления линейной регрессионной зависимости между факторами. Установление значимости регрессионной модели.

Используемые программные средства. MS Excel 2010 (2016), STATISTICA 8.

6.1. Краткие теоретические сведения.

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 $X_1, X_2, \dots, X_k$ (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.

$$\varepsilon_i \sim N\ 0, \sigma\ ; \quad i = 1, n$$

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:

$$S = \sum_{i=1}^{n} \left(y_i - (a_0 + a_1 x_i) \right)^2 \to min$$

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:

$$a_1 = \frac{K_{xy}}{s_x^2} = r_B \frac{s_y}{s_x}; \;\; a_0 = y - a_1 x \qquad (6.1)$$

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

Уравнение вида $y_x = a_0 + a_1 x$ называется *уравнением линейной регрессии* Y на X, а получаемые из него значения $y_i = a_0 + a_1 x_i$ 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 6.1. Аналогично можно найти выборочное уравнение прямой линии регрессии X на Y:

$$x_y = \rho_{x/y} y + c$$

где $\rho_{x/y}$ – выборочный коэффициент регрессии X на Y.

В формуле (6.1) K_{xy} и r_B есть соответственно эмпирические *корреляционный момент (ковариация)* и *коэффициент корреляции* для величин X и Y.

Замечание 6.2. 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:

$$y_x - y = r_B \frac{s_y}{s_x} (x - x) \text{ и } x_y - x = r_B \frac{s_x}{s_y} (y - y).$$

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

$$D_{\text{ост}} = \frac{1}{n-2} \sum_{i=1}^{n} \left(y_i - y\ x_i \right)^2 = \frac{Q_e}{n-2} \qquad (6.2)$$

Величина Q_e называется *остаточной суммой квадратов* (или *суммарной невязкой*). Разности между наблюдаемыми значениями переменной Y при $x = x_i$ и расчётными значениями $y_x\ x_i = a_0 + a_1 x_i$ 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 $Q_R = \sum_{i=1}^{n} \left(y_x\ x_i - y \right)^2$.

Тогда выборочное значение F, имеющее распределение Фишера:

$$F = \frac{Q_R\ n-2}{Q_e} \qquad (6.3)$$

Уравнение регрессии значимо, если $F_{\text{набл}} > F_{\text{крит}}\ \alpha, 1, n-2$ 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.

Нелинейная регрессия.

Ввиду простоты расчетов линейная зависимость используется довольно часто. Кроме того, многие функции, зависящие от двух параметров, можно линеаризовать путем *замены переменных*.

Для этого необходимо подобрать такое преобразование исходной зависимости $y\ x = \varphi\ a_0, a_1, x$, в результате которого зависимость приобретает линейный вид $v = b_0 + b_1 u$. Далее решается задача линейной аппроксимации для новой зависимости и вычисленные коэффициенты b_0 и b_1 пересчитываются в коэффициенты a_0 и a_1.

Для ряда часто встречающихся двухпараметрических зависимостей возможные замены переменных приведены в табл. 6.1.

Таблица 6.1

№	Вид зависимости	Уравнение регрессии	Сведение к линейному виду	Обратная замена переменных
1	Гиперболическая	$y_x = a_0 + \frac{a_1}{x}$	$v = y; u = \frac{1}{x}$ $x \neq 0$	$a_0 = b_0, a_1 = b_1$
2	Логарифмическая	$y_x = a_0 + a_1 \ln x$	$v = y; u = lnx$ $x > 0$	$a_0 = b_0, a_1 = b_1$
3	Показательная	$y_x = a_0 + a_1 e^x$	$v = y; u = e^x$	$a_0 = b_0, a_1 = b_1$
		$y_x = a_0 e^{a_1 x}$	$v = lny; u = x;$ $y > 0;\ a_0 > 0$	$a_0 = e^{b_0}, a_1 = b_1$
4	Степенная	$y_x = a_0 x^{a_1}$	$v = lny;$ $u = lnx;$ $x, y > 0;\ a_0 > 0$	$a_0 = e^{b_0}, a_1 = b_1$
5	Параболическая	$y_x = a_0 + a_1 \ \overline{x}$	$v = y; u = \ \overline{x}$ $x \geq 0$	$a_0 = b_0, a_1 = b_1$
6	Параболическая	$y_x = a_0 + a_1 x + a_2 x^2$	К линейной не сводится	

В случаях 1–5 параметры линейной зависимости находятся по формулам (6.1). Для случая 6 применяется непосредственно метод наименьших квадратов, после применения которого оценки параметров модели находятся из системы

$$\begin{matrix} a_0 n + a_1 \ x_i + a_2 \ x_i^2 = \ y_i \\ a_0 \ x_i + a_1 \ x_i^2 + a_2 \ x_i^3 = \ x_i y_i \\ a_0 \ x_i^2 + a_1 \ x_i^2 + a_2 \ x_i^4 = \ x_i^2 y_i \end{matrix} \quad (6.4)$$

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

Качество регрессионной модели оценивается с помощью коэффициента детерминации, который вычисляется по формуле:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} y_i - y_i \ ^2}{\sum_{i=1}^{n} y_i - y \ ^2}, \quad (6.5)$$

где y_i – фактическое значение зависимой величины в i –м наблюдении; $y_i = y \ x_{1i}, x_{2i}, \dots, x_{ni}$ – значение зависимой переменной, определяемой по уравнению регрессии; $y = \frac{1}{n} \ _i y_i$ – среднее арифметическое фактических значений зависимой переменной.

Для небольших значений $n < 30$ необходимо использовать скорректированный коэффициент детерминации:

$$R^{*2} = 1 - \frac{n-1}{n-m}\ 1 - R^2\ .$$

Здесь m – число оцениваемых параметров.

Замечание 6.3. 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 более 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.

6.2. Практическая часть.

Контрольный пример 6.1. При исследовании зависимости между случайными величинами X и Y была получена следующая таблица измерений соответствующих значений этих величин (см. лабораторную работу № 5):

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. Аппроксимировать статистическую зависимость между этими величинами линейной функцией $y = a_1 x + a_0$; проверить модель на значимость (адекватность). Найти остаточную дисперсию и $\sigma_{\text{ост}}$.
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. 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 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. 6.3): имеем основные результаты: скорректированный коэффициент детерминации R^2: 0.98578061; гипотеза о нулевом значении наклона отклоняется с высоким уровнем значимости $p = 0.000001$.

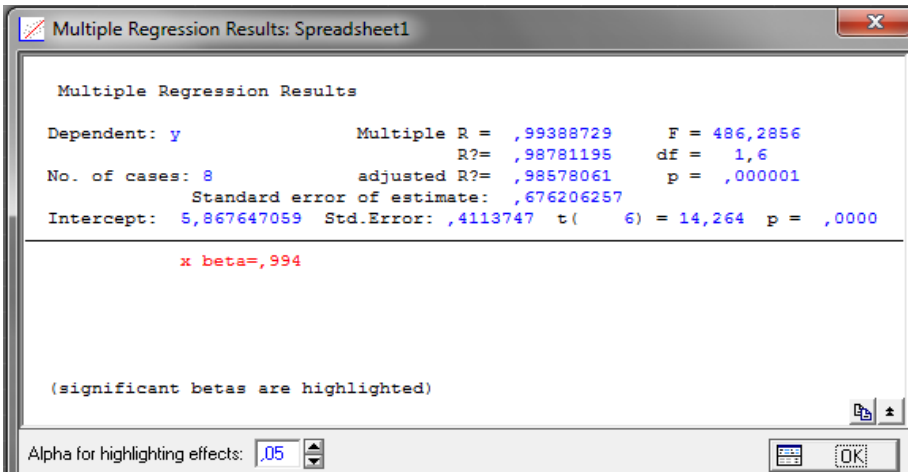


Рис. 6.2 – Окно результатов регрессионного анализа

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. 6.3):

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. 6.3 – 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. 6.4):

Summary Statistics; DV: y (Spreadsheet1)

Statistic	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. 6.4 – Вкладка *Summary Statistics*

Так как стандартная ошибка оценки (*Std Error of estimate)* – небольшая, то наблюдаемые значения близки к предсказываемым. Значение коэффициента детерминации $R^{*2} = \text{Adjusted } R? = 0.9858$ 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*.

Для оценки адекватности модели построим нормальный вероятностный график остатков.

В отобразившемся окне, перейдя к вкладке *Quick*, необходимо нажать кнопку *Normal plot of residuals*. Полученный график остатков приведён на рис. 6.5.

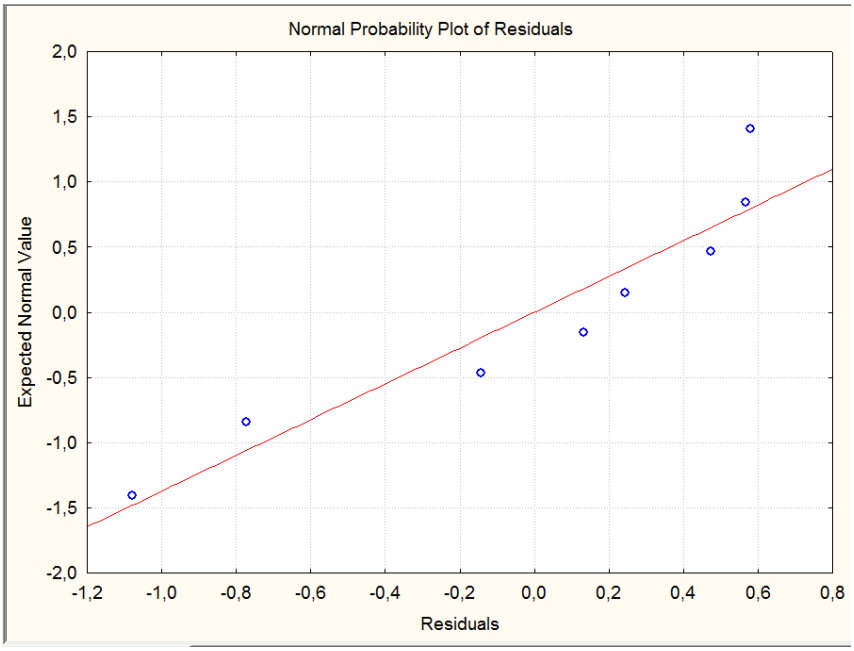


Рис. 6.5 – Нормальный вероятностный график остатков

Из этого графика видно, что остатки хаотично разбросаны относительно прямой, в их поведении нет закономерностей. Нет оснований говорить, что остатки связаны между собой. Отсюда можно заключить, что модель достаточно адекватно описывает данные.

Так как мы имеем очень небольшое число данных – 8, поэтому используются графические методы оценки адекватности модели. В сложных задачах графические и статистические методы оценки адекватности должны естественно дополнять друг друга. Продемонстрируем это, вернувшись в окно *Residual Analysis*, кнопкой активизировав окно *Summary: Residual&Predicted*:

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. 6.6. Наблюдаемые и предсказанные значения остатков

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. 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. 6.7 (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. 6.7).

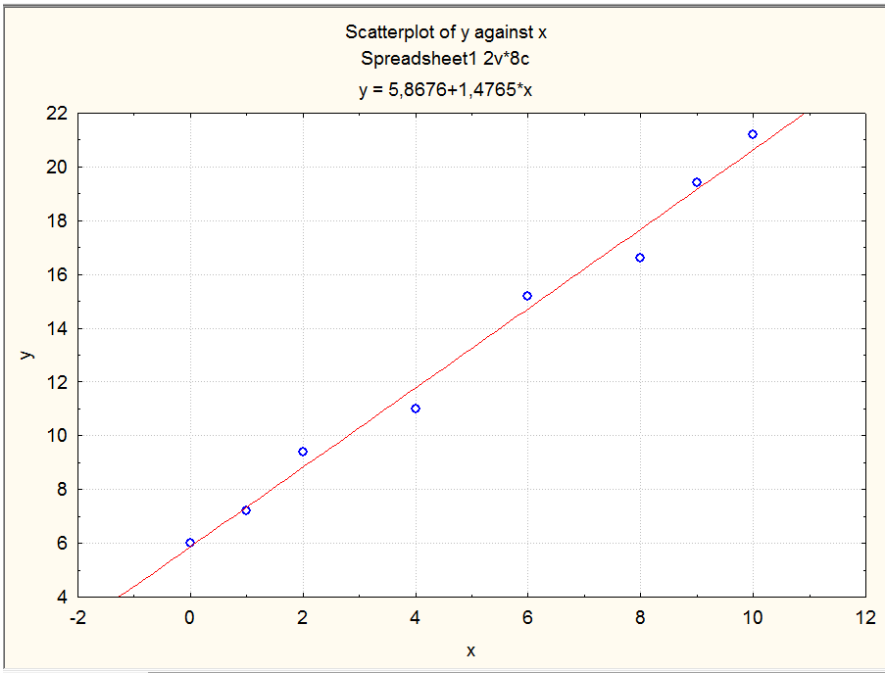


Рис. 6.7 – Диаграмма рассеяния с подобранной прямой линией регрессии

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. 6.8, диапазон A2: B9). Далее откроем меню инструмента «*Анализ данных*». Выбираем инструмент «*Регрессия*». Заполним диалоговое окно, как показано на рис. 6.8.

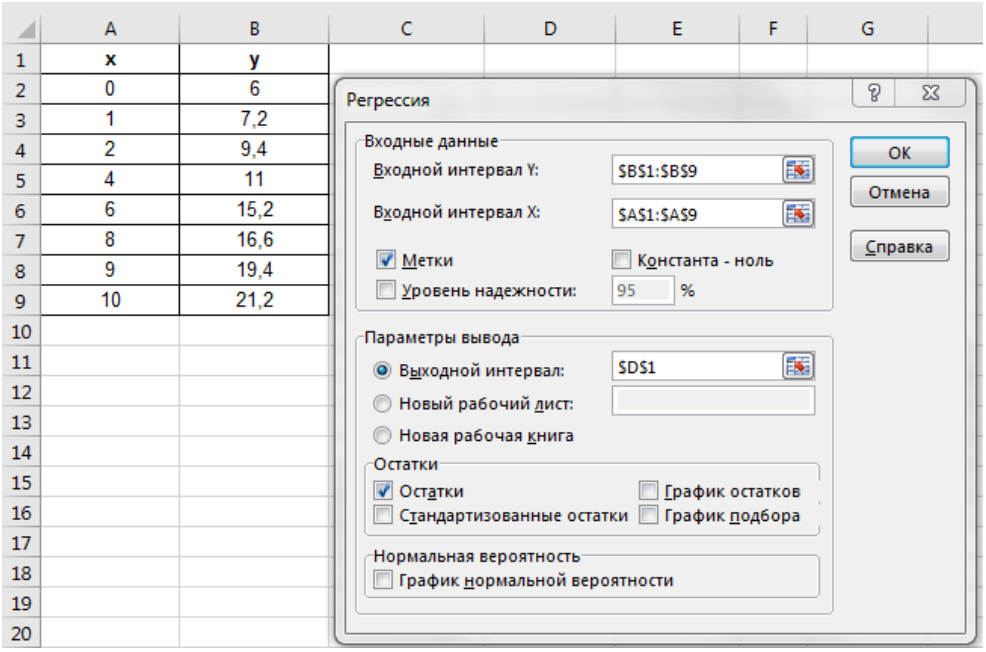


Рис. 6.8 – Исходные данные задачи и диалоговое окно инструмента «Регрессия»

После нажатия ОК, программа отобразит расчеты:

	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. 6.9 Результаты анализа линейной модели регрессии

Используя оценки $a_0 = 5{,}8676$ и $a_1 = 1{,}4765$ (ячейки E17 и E18) параметров регрессии a_0 и a_1, запишем выборочное уравнение парной линейной регрессии:

$$y_x = 5{,}8676 + 1{,}4765x.$$

Оценим с помощью средней ошибки аппроксимации качество уравнения. Воспользуемся результатами регрессионного анализа, представленного на рис. 6.10.

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. 6.10 – Результат применения инструмента «Регрессия» (Вывод остатка)

Составим новую таблицу, как показано на рис. 6.11. В столбце E рассчитаем относительную ошибку аппроксимации по формуле:

$$\mathrm{A_i} = \frac{y - y_x}{y} \cdot 100$$

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. 6.11 Расчет средней ошибки аппроксимации

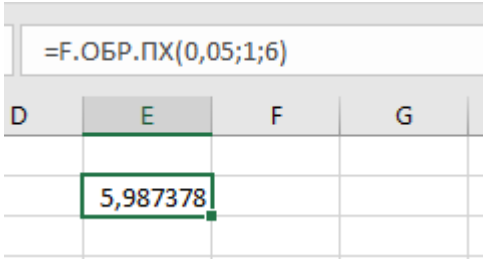
Средняя ошибка аппроксимации рассчитывается по формуле:

$$\mathrm{A} = \frac{1}{n} \qquad \mathrm{A_i} = \frac{30{,}95728}{8} \approx 3{,}9.$$

Качество построенной модели оценивается как хорошее, так как А не превышает 8%.

Выборочный коэффициент корреляции $r_B = 0{,}9939 > 0{,}7$, 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.

Найдём критическое значение F-критерия при уровне значимости 0,05 и числе степеней свободы $k_1 = m - 1 = 2 - 1 = 1$ и $k_2 = n - m = 8 - 2 = 6$, используя стандартную функцию $F.\text{ОБР.ПХ}(0{,}05;\, 1;\, 6)$ пакета Excel:

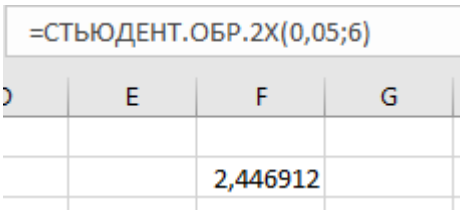


Так как рассчитанное значение критерия больше табличного ($F =$

$486.2856 > 5{,}987$), то уравнение регрессии признаётся *значимым*.

Оценку статистической значимости параметров регрессии проведём с помощью t-статистики Стьюдента и путём расчёта доверительного интервала каждого из показателей.

Выдвигаем гипотезу H_0 о статистически незначимом отличии показателей от нуля: $H_0: a_0 = a_1 = r_{XY} = 0$. $t_{крит} = 2{,}447$ для числа степеней свободы $df = n - 2 = 8 - 2 = 6$ и $\alpha = 0{,}05$:



На рисунке 6.9 имеются фактические значения t-статистики:

$$t_{a_0} = 14{,}2635;\ t_{a_1} = 22{,}0519.$$

t-критерий для коэффициента корреляции можно рассчитать по формуле:

$$t_r = \overline{F} = \overline{486{,}2856} \approx 22{,}1.$$

Фактические значения t-статистики превосходят табличные значения:

$$t_{a_0} = 14{,}2635 > t_{крит} = 2{,}447;$$
$$t_{a_1} = 22{,}0519 > t_{крит} = 2{,}447;$$
$$t_r = 22{,}1 > t_{крит} = 2{,}447.$$

Поэтому гипотеза H_0 отклоняется, то есть параметры регрессии и коэффициент корреляции не случайно отличаются от нуля, а статистически значимы.

Для параметра a_0 95%-ные границы (как показано на рис. 6.9) составили:

$$4{,}861 \le a_0 \le 6{,}8742.$$

Для коэффициента регрессии a_1 95%-ные границы (как показано на рис. 6.9) составили :

$$1{,}3126 \le a_1 \le 1{,}6403$$

Анализ верхней и нижней границ доверительных интервалов 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.

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.

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*, которая имеет вид, представленный на рис. 6.12.

Выберем тип линии тренда – *Линейная* и установим флажки:

- *показывать уравнение на диаграмме*;
- *поместить на диаграмму величину достоверности аппроксимации (R^2)*.

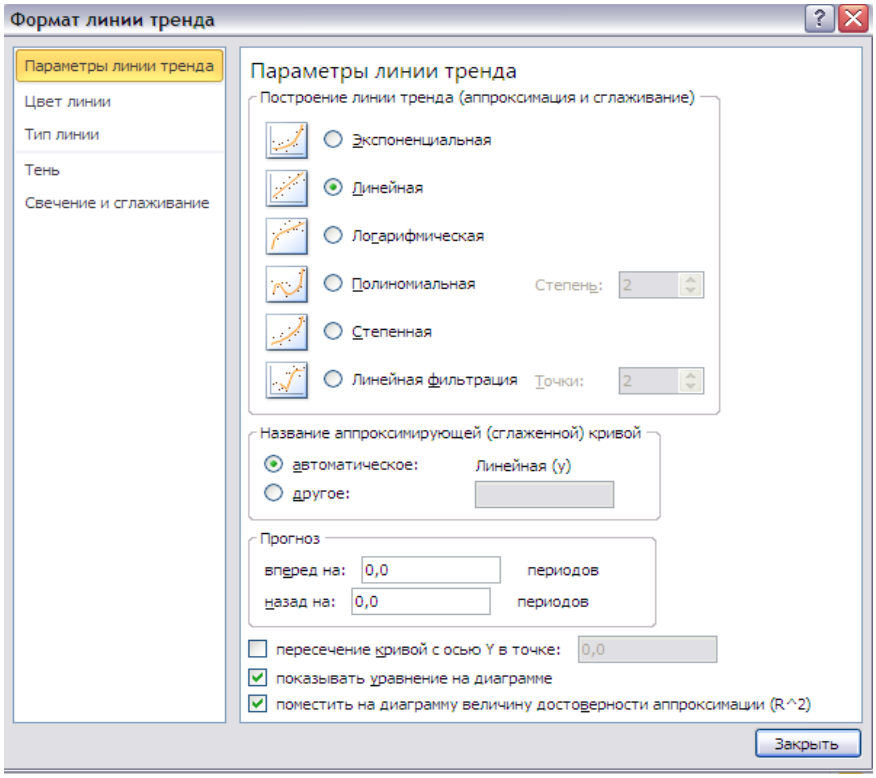


Рис. 6.12 – Вид диалогового окна «Формат линии тренда»

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. 6.13).

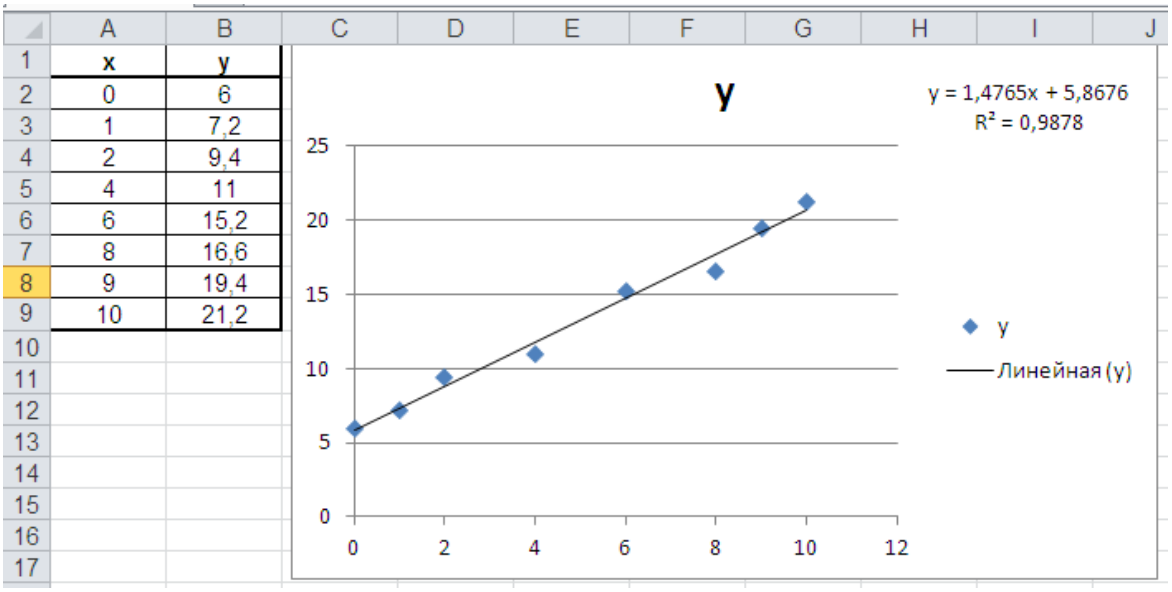


Рис. 6.13 – Вид рабочего листа Excel. Линейный тренд.

Контрольный пример 6.2. Найти оценки параметров модели $y_x = a_0 + a_1 x + a_2 x^2$ по следующим данным:

x	-3	-2	-1	0	1	2	3
y	-10	0	4	5	4	2	-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*.

Решение в пакете Excel. Решение задачи можно разбить на следующие этапы: ввод исходных данных, построение точечного графика и добавление к этому графику линии тренда.

Введем исходные данные в рабочий лист и построим график экспериментальных данных. Далее выделим экспериментальные точки на графике, щелкнем правой кнопкой мыши и воспользуемся командой «*Добавить линию тренда*». Выберем в качестве аппроксимирующей зависимости полином второй степени и выведем уравнение, описывающее этот полином на график. Полученная диаграмма представлена на рис. 6.14.

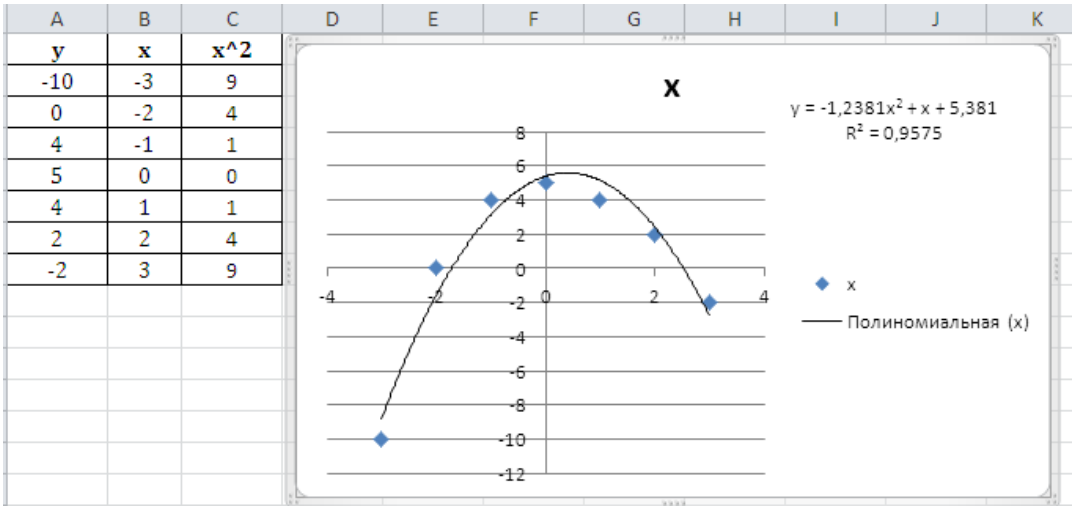


Рис. 6.14 – Исходные данные и графическая иллюстрация результатов анализа модели регрессии $y_x = a_0 + a_1 x + a_2 x^2$

Подберём коэффициенты заданной зависимости с помощью статистической процедуры *Регрессия*.

Воспользуемся командой *Данные – Анализ данных***.** В открывшемся окне выделим процедуру *Регрессия* и щёлкнем на кнопке ОК. Заполним диалоговое окно процедуры, как показано на рисунке 6.15.

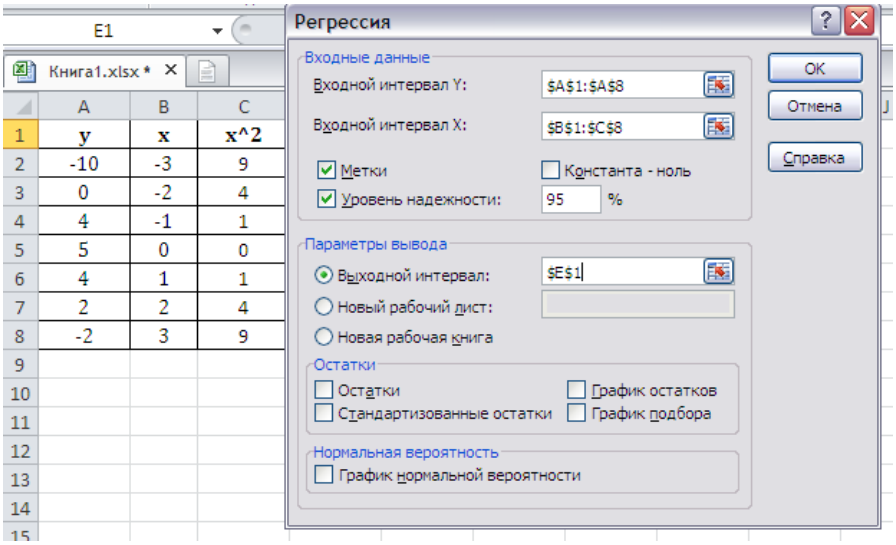


Рис. 6.15 – Диалоговое окно процедуры *Регрессия*

Щелчком на кнопке ОК запустим процедуру *Регрессия.*

На данном рабочем листе появятся три таблицы результатов реализации процедуры (рис. 6.16):

	E	F	G	H	I	J	K
1	ВЫВОД ИТОГОВ						
2							
3	*Регрессионная статистика*						
4	Множественный R	0,979					
5	R-квадрат	0,958					
6	Нормированный R-квадрат	0,936					
7	Стандартная ошибка	1,318					
8	Наблюдения	7,000					
9							
10	Дисперсионный анализ						
11		*df*	*SS*	*MS*	*F*	*Значимость F*	
12	Регрессия	2	156,762	78,381	45,096	1,803E-03	
13	Остаток	4	6,952	1,738			
14	Итого	6	163,714				
15							
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*
17	Y-пересечение	5,381	0,761	7,069	0,002	3,268	7,494
18	x	1,000	0,249	4,014	0,016	0,308	1,692
19	x^2	-1,238	0,144	-8,607	0,001	-1,637	-0,839

Рис. 6.16 – Результаты анализа полиномиальной модели регрессии

$$y_x = a_0 + a_1 x + a_2 x^2$$

Использовав оценки $a_0 = 5{,}381;\ a_1 = 1;\ a_2 \approx -1{,}238$ (ячейки $F17, F18, F19$) параметров регрессии $a_0,\ a_1, a_2$, запишем выборочное уравнение полиномиальной регрессии $y_x = 5{,}382 + x - 1{,}238\,x^2$.

Близкий к единице коэффициент детерминации (ячейка F5), большое расчётное значение статистики F (ячейка I12) и малая значимость F (ячейка J12) свидетельствуют о высокой адекватности полиномиальной модели.

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 $\gamma = 0{,}95$. Ни один из этих интервалов не накрывает нуль.

Решим данную задачу с применением пакета *Statistica*. Вводим исходные данные для переменных x и y (рис. 6.17):

	1 x	2 y
1	-3	-10
2	-2	0
3	-1	4
4	0	5
5	1	4
6	2	2
7	3	-2

Рис 6.17 – Ввод исходных данных

Проведем анализ в модуле *Nonlinear estimation (Нелинейная оценка)* (рис. 6.18)

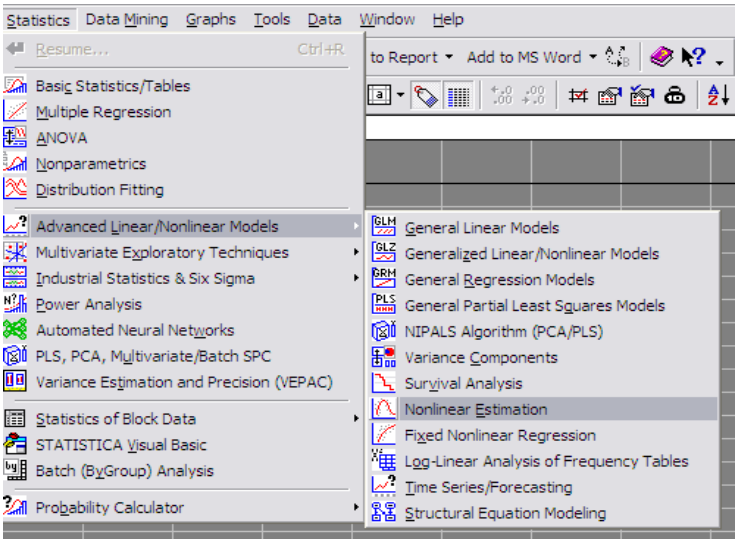


Рис. 6.18 – Запуск модуля *Nonlinear estimation*

На экране появится стартовая панель модуля. Выберем опцию *User specified regression, least squares* (Метод наименьших квадратов) и щелкнем мышью по названию модуля.

В появившемся окне щелкнем мышью по кнопке *Function of estimated* (*Предполагаемая функция*).

В окне с клавиатуры введем предполагаемую функцию. В пакете *Statistica* можно ввести любую формулу, связывающую зависимую и независимую переменные.

В данном случае предполагается, что наиболее подходящей функцией является полином второй степени типа $Y = a_0 + a_1X + a_2X^2$ или в конкретном случае в соответствии с таблицей $y = a0 + a1 * x + a2 * x ** 2$ (рис. 6.19):

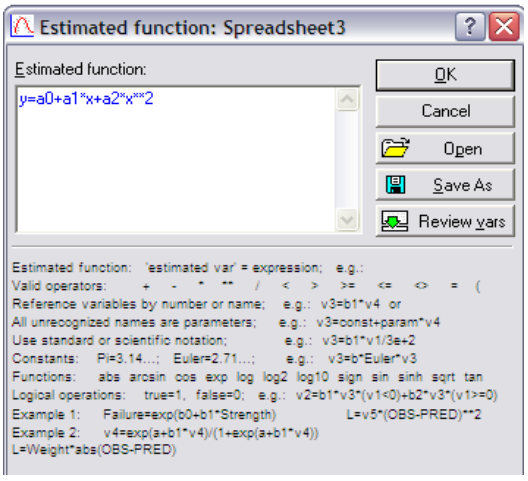


Рис. 6.19 – Ввод функции

В нижней части рисунка приведен перечень алгебраических и функ-

циональных символов, которые воспринимаются программой.

Нажимаем ОК. Затем еще раз ОК.

Верхняя часть окна информирует о модели, методе, количестве взятых в анализ пар. В середине окна выберем метод аппроксимации. Нажмем ОК.

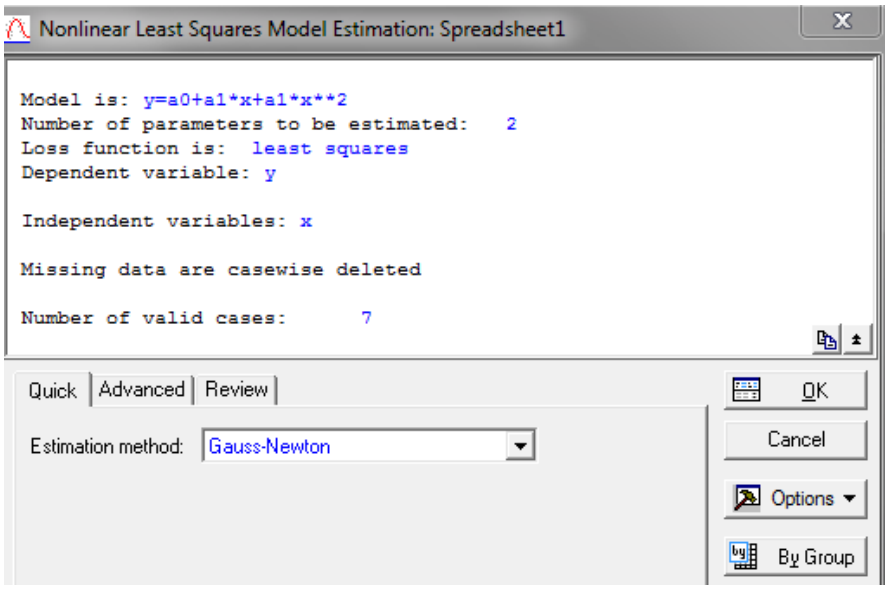


Рис. 6.20 Панель пуска аппроксимации

В верхней части появившегося окна результатов (рис. 6.21) показаны значения корреляционного отношения $\eta_{y/x} = \overline{R^2} = 0{,}97853637$ и его квадрата $R^2 = 0{,}95753345$. Это указывает на сильную корреляционную связь между переменными.

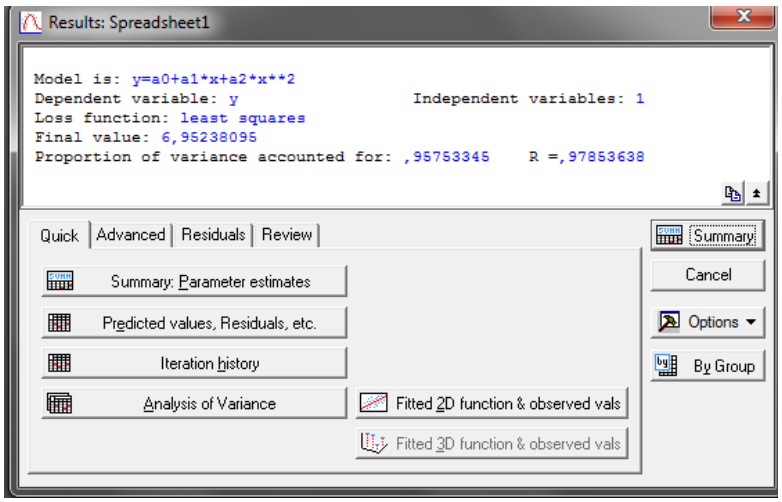


Рис. 6.21 – Окно результатов

В окне результатов щелкнем мышью по кнопке *Summary: Parameter estimates* (Итоговые параметры и стандартные ошибки). Полученные результаты представлены на рис. 6.22.

Model is: y=a0+a1*x+a2*x**2 (Spreadsheet1)
Dep. Var. : y
Level of confidence: 95.0% (alpha=0.050)

	Estimate	Standard error	t-value df = 4	p-level	Lo. Conf Limit	Up. Conf Limit
a0	5,38095	0,761160	7,06941	0,002113	3,26763	7,494272
a1	1,00000	0,249148	4,01368	0,015948	0,30825	1,691746
a2	-1,23810	0,143846	-8,60710	0,001001	-1,63748	-0,838715

Рис. 6.22 – Результаты аппроксимации

Аппроксимирующая функция: $y_x = 5{,}381 + x - 1{,}2381x^2$.

В столбце *Estimate* (Оценка) показаны значения коэффициентов. Далее указаны стандартные ошибки, *t-критерий* при 4 степенях свободы, уровень значимости, верхний и нижний пределы доверительных интервалов для коэффициентов регрессии.

Так как доверительные интервалы для коэффициентов a_0, a_1 и a_2 не содержат нулевое значение, то эти коэффициенты значимо (существенно) отличаются от нуля. Этот вывод подтверждает и выделение пакетом красным цветом строк, соответствующих этим коэффициентам, а также низкие (меньше уровня значимости 0,05) значения значимости p.

В окне результатов (рис. 6.21) в режиме *Quick* нажмем кнопку *Analysis of Variance* (Дисперсионный анализ). Результат выполненной операции представлен на рис. 6.23, который свидетельствует о достоверности регрессии ($F = 30{,}3105$ при $p = 0{,}003282$).

Model is: y=a0+a1*x+a2*x**2 (Spreadsheet1)
Dep. Var. : y

Effect	1 Sum of Squares	2 DF	3 Mean Squares	4 F-value	5 p-value
Regression	158,0476	3,000000	52,68254	30,31050	0,003282
Residual	6,9524	4,000000	1,73810		
Total	165,0000	7,000000			
Corrected Total	163,7143	6,000000			
Regression vs.Corrected Total	158,0476	3,000000	52,68254	1,93077	0,225904

Рис. 6.23 – Результат дисперсионного анализа

В окне рис.6.21 перейдем в режим просмотра результатов *Residials* (Остатки). Щелкнем мышью по кнопке *Observed, predicted, residual vals* (Наблюдаемый, предсказанный, остаточный). Результаты выполненной операции представлены на рис. 6.24.

Model is: y=a0+a1*x+a2*x**2 (Spreadsheet1)
Dep. Var. : y

	Observed	Predicted	Residuals
1	-10,0000	-8,76190	-1,23810
2	0,0000	-1,57143	1,57143
3	4,0000	3,14286	0,85714
4	5,0000	5,38095	-0,38095
5	4,0000	5,14286	-1,14286
6	2,0000	2,42857	-0,42857
7	-2,0000	-2,76190	0,76190

Рис. 6.24 – Наблюдаемые и аппроксимированные значения функции

Построим корреляционное поле и линию регрессии на корреляционном поле. Для этого в меню *Graphs* выберем команду *Scatter plots*. В полученном окне нажмем кнопку *Variables*, и установим зависимые данные – x, y и опции графика – *Graphs Type: Regular, Fit(подбор): Polynomial*. Наблюдаем диаграмму рассеяния с подобранной кривой, параметры которой отражены в ее заголовке (рис. 6.25).

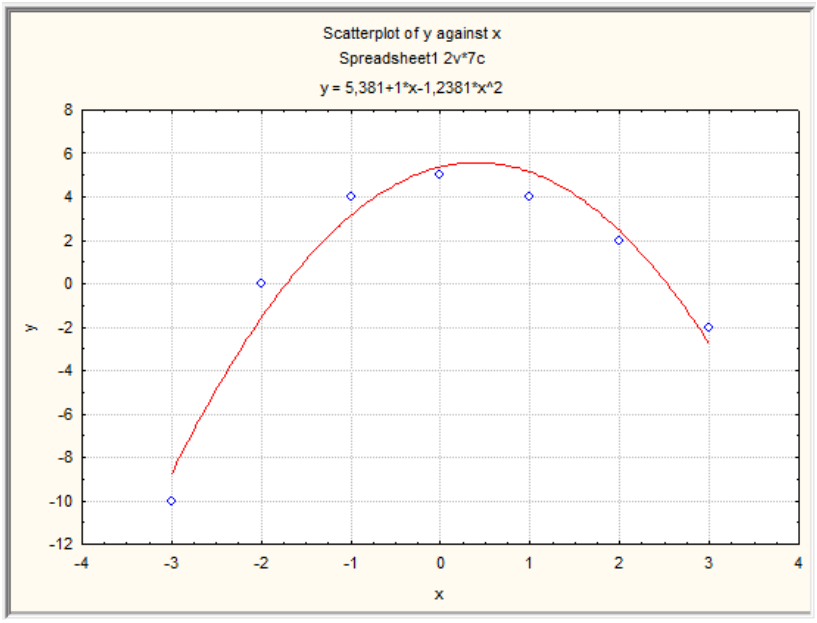


Рис. 6.26 – Диаграмма рассеяния с подобранной кривой

6.3. Задания для самостоятельной работы.

Задание 1. Используя данные своего варианта из лабораторной работы № 5 (задание 1):

1) Аппроксимировать статистическую зависимость между этими величинами линейной функцией $y = a_1 x + a_0$; проверить модель на значимость (адекватность). Найти остаточную дисперсию и $\sigma_{\text{ост}}$.
2) Вычислить коэффициент детерминации, сделать вывод.
3) Построить корреляционное поле и линию регрессии на корреляционном поле. В пакете *Statistica* провести анализ остатков.
   Принять $\alpha = 0{,}05$.

Задание 2. Используя данные своего варианта из лабораторной работы № 5 (задание 3):

1. построить диаграмму рассеивания;
2. по виду полученной диаграммы подобрать 3-4 типа функциональных зависимостей (см. табл. 6.1);
3. провести аппроксимацию методом наименьших квадратов;
4. оценить результаты аппроксимации – для каждого из полученных эмпирических формул вычислить коэффициент детерминации. Выбрать эмпирическую формулу, которая более точно описывает результаты эксперимента.

Лабораторная работа № 7.

НЕПАРАМЕТРИЧЕСКИЕ МЕТОДЫ МАТЕМАТИЧЕСКОЙ СТАТИСТИКИ

Цель работы: приобретение практических навыков в построении статистики и проверке непараметрической статистической гипотезы при неизвестном законе распределения случайной величины.

Используемые программные средства: STATISTICA 8.0.

7.1. Краткие теоретические сведения

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 сами 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.

Для каждого параметрического критерия имеется, по крайней мере, один непараметрический аналог. Эти критерии можно отнести к одной из следующих групп:

- критерии различия между группами (независимые выборки);
- критерии различия между группами (зависимые выборки);
- критерии зависимости между переменными.

*Различия между независимыми группами***.** Обычно, когда имеются две выборки, которые нужно хотите сравнить относительно среднего значения некоторой изучаемой переменной, используется *t-критерий для независимых выборок*. Непараметрическими альтернативами этому критерию являются: критерий *серий Вальда-Вольфовица*, *U* критерий *Манна-Уитни* и *двухвыборочный критерий Колмогорова-Смирнова*. Для нескольких групп используют *дисперсионный анализ*. Его непараметрическими аналогами являются: ранговый дисперсионный анализ *Краскела-Уоллиса* и *медианный* тест.

*Различия между зависимыми группами***.** При сравнении двух переменных, относящихся к одной и той же выборке (например, математические успехи студентов в начале и в конце семестра), то обычно используется *t-критерий для зависимых выборок*. Альтернативными непараметрическими тестами являются: критерий *знаков* и критерий *Вилкоксона парных сравнений*. Если рассматриваемые переменные по природе своей категориальны

или являются категоризованными (т.е. представлены в виде частот попавших в определенные категории), то подходящим будет критерий *хи-квадрат Макнемара.* Если рассматривается более двух переменных, относящихся к одной и той же выборке, то обычно используется дисперсионный анализ (ANOVA) с повторными измерениями. Альтернативным непараметрическим методом является *ранговый дисперсионный анализ Фридмана* или Q критерий *Кохрена* (последний применяется, например, если переменная измерена в номинальной шкале). Q критерий Кохрена используется также для оценки изменений частот (долей).

*Зависимости между переменными***.** Для того чтобы оценить зависимость (связь) между двумя переменными, обычно вычисляют коэффициент корреляции. Непараметрическими аналогами стандартного коэффициента корреляции Пирсона являются статистики *Спирмена* ρ, τ *Кендалла* (см. лабораторную работу № 4) и коэффициент *Гамма*. Если две рассматриваемые переменные по природе своей категориальны, подходящими непараметрическими критериями для тестирования зависимости будут: *Хи-квадрат, Фи* коэффициент, *точный критерий Фишера*. Дополнительно доступен критерий зависимости между несколькими переменными так называемый *коэффициент конкордации Кендалла*. Этот тест часто используется для оценки согласованности мнений независимых экспертов (судей), в частности, баллов, выставленных одному и тому же субъекту.

Большая группа непараметрических критериев используется для проверки гипотезы о проверки двух выборок $x_1, \dots, x_n$ и $y_1, \dots, y_n$ одной и той же генеральной совокупности, то есть о том, что функции распределения $F_X(x)$ и $F_Y(y)$ двух генеральных совокупностей равны: $F_X\ x\ \equiv F_Y\ y$. 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.

При решении конкретной задачи необходимо выбрать тот или иной метод. *Первым критерием* для выбора метода является вид шкалы, в которой представлены исходные данные. *Вторым критерием* является вид выборок (независимые или связанные) и их количество.

Связанные (зависимые) выборки характеризуются тем, что измерения проводятся на одной и той же группе, состоящей из n объектов, находящихся в различных условиях. В случае если каждый из n объектов подвергается m воздействиям, то результаты наблюдений представляют m связанных выборок объема n. В связанных выборках количество наблюдений одинаково. Независимые (несвязанные) выборки такими свойствами не обладают.

Таким образом, рассматриваемые ниже методы можно квалифицировать следующим образом:

1) *Исходные данные*: две независимые выборки объемов n_1 и n_2. *Проверяемая гипотеза* H_0: выборки принадлежат однородным генеральным совокупностям.

Методы:

- критерий серий Вальда-Вольфовица;
- критерий Манна-Уитни;
- двухвыборочный критерий Колмогорова-Смирнова.

2) *Исходные данные*: m независимых выборок, объемы которых соответственно равны $n_1, n_2, \ldots, n_m$. *Проверяемая гипотеза* H_0: выборки принадлежат однородным генеральным совокупностям.

Методы:

- однофакторный дисперсионный анализ Краскелла-Уоллиса;
- медианный критерий.

3) *Исходные данные*: две связанные выборки объема n. *Проверяемая гипотеза* H_0: выборки принадлежат однородным генеральным совокупностям.

Методы:

- критерий знаков;
- знако-ранговый критерий Вилкоксона.

4) *Исходные данные*: m связанных выборок объема n. *Проверяемая гипотеза* H_0: выборки принадлежат однородным генеральным совокупностям.

Методы: двухфакторный анализ Фридмана;

Меры связи: коэффициент конкордации Кендалла.

Непараметрические методы наиболее приемлемы, когда объем выборок мал. Если данных много (например, $n > 100$), то не имеет смысла использовать непараметрические статистики.

7.2. Практическая часть.

Контрольный пример 7.1. 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

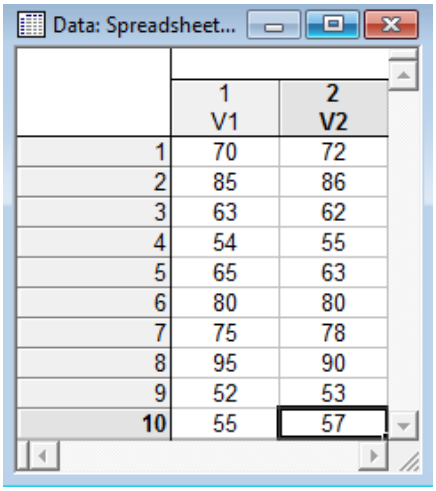
Решение. Критерий знаков является непараметрической альтернативой *t*-критерию Стьдента в случае зависимых выборок, который применяется, когда

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:



	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. 7.1 – Таблица исходных данных

Для запуска модуля *Непараметрические статистики* в меню *Statistics* необходимо выбрать *Nonparametrics*, стартовая панель изображена на рис. 7.2

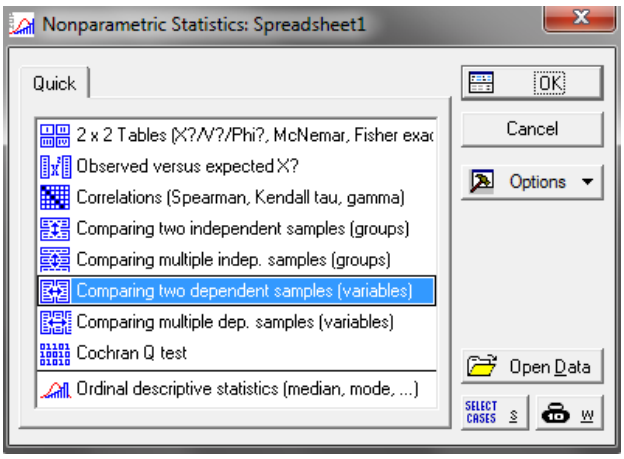


Рис. 7.2 – Стартовая панель модуля *Nonparametrics*

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*):

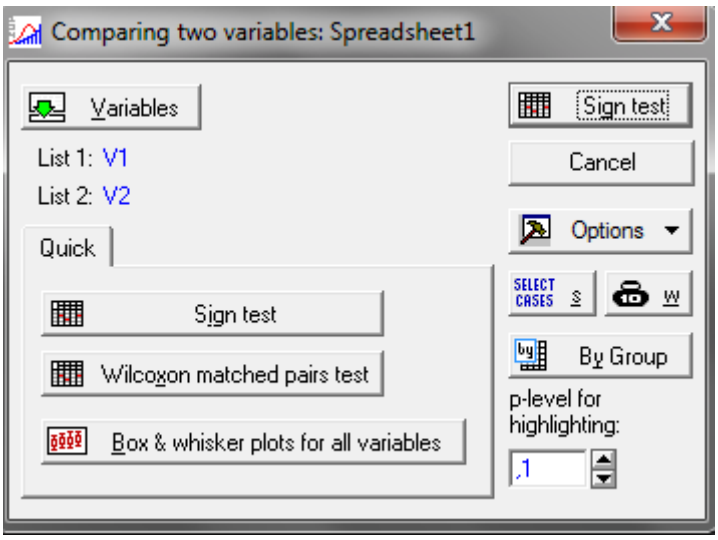


Рис. 7.3 – Задание переменных

Нажав на кнопку *Sign test*, рассчитаем характеристики для критерия знаков:

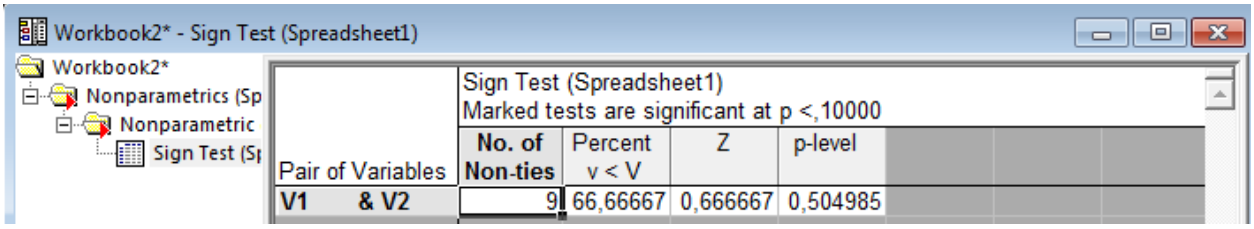
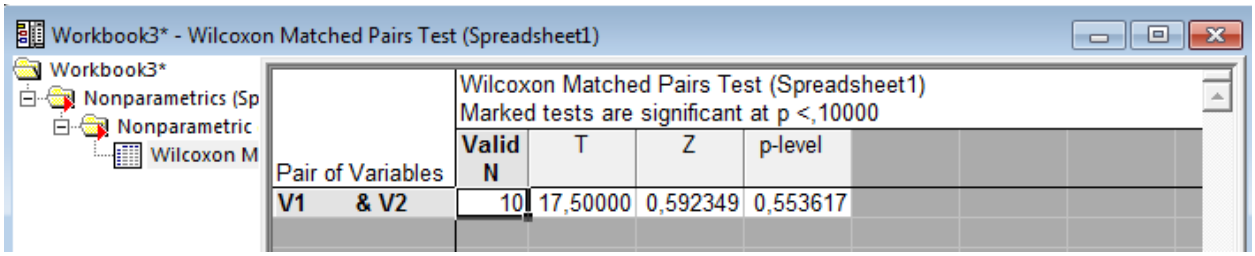


Рис. 7.4 – Результаты критерия знаков

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. 7.5 – Результаты критерия Вилкоксона

Как видно из рис. 7.5, 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.7.3.

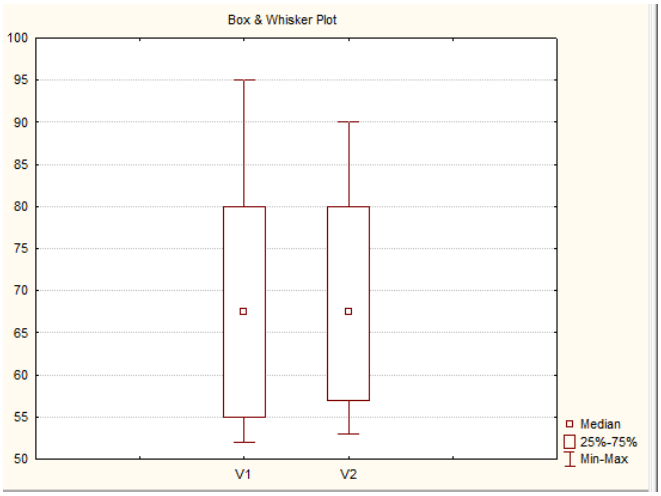


Рис. 7.6 – Диаграмма размаха

На диаграмме размаха для каждой переменной показаны: медиана, квартильный размах (25% и 75%), размах (минимум, максимум).

Контрольный пример 7.2. 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. 7.7):

	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. 7.7 – 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.

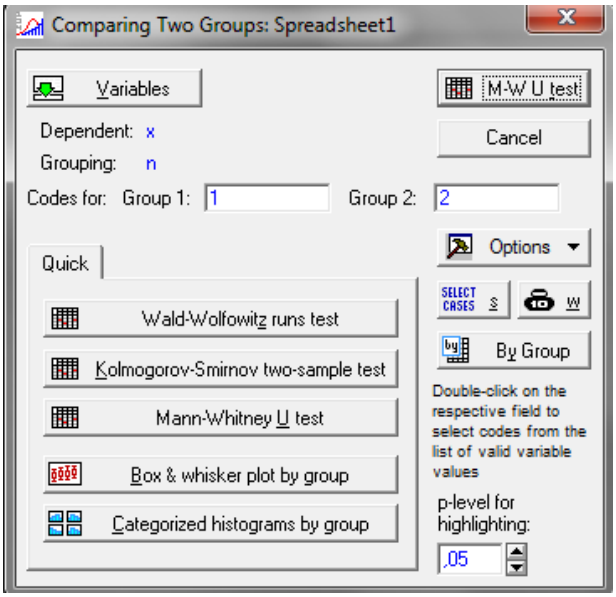
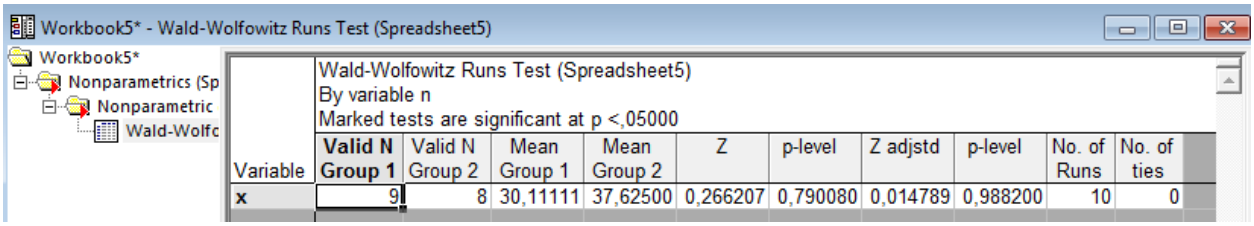


Рис. 7.8 – Задание зависимой и группирующей переменных

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.



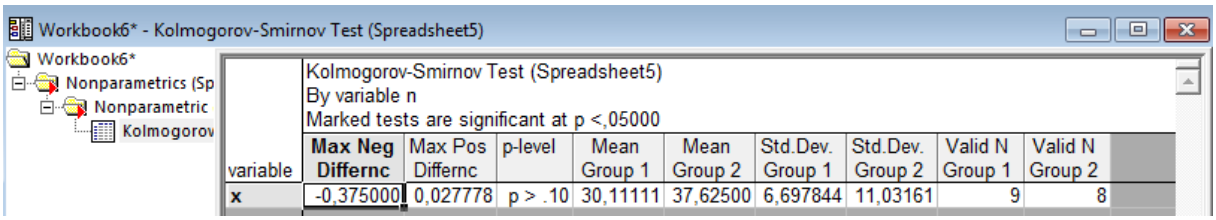
Wald-Wolfowitz Runs Test (Spreadsheet5)
By variable n
Marked tests are significant at p <,05000

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. 7.9 – 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. 7.10 – 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.

Можно заметить, что стандартные отклонения в обеих группах не равны (см. рис. 7.9 и рис. 7.10), следовательно, невозможно применить *t*-критерий.

Mann-Whitney U Test (Spreadsheet1)
By variable n
Marked tests are significant at p <,05000

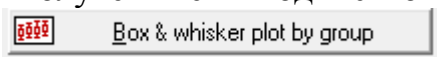
variable	Rank Sum Group 1	Rank Sum Group 2	U	Z	p-level	Z adjusted	p-level	Valid N Group 1	Valid N Group 2	2*1sided exact p
x	68,00000	85,00000	23,00000	-1,25093	0,210963	-1,25246	0,210403	9	8	0,235870

Рис. 7.11 – 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. 7.12) – кнопка 

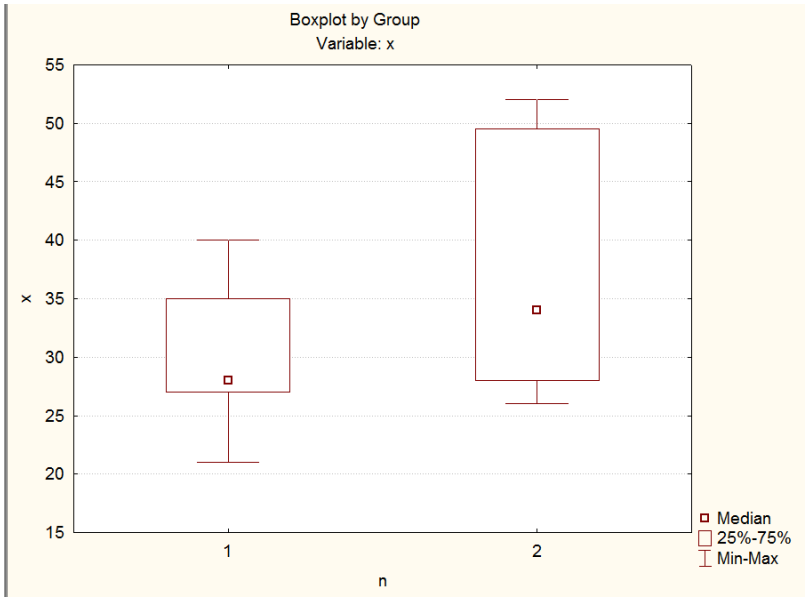


Рис. 7.12 – Диаграмма размаха

Контрольный пример 7.3. 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. 7.13):

	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. 7.13 – Исходная выборка данных
(Error – ошибки; Code – код)

В стартовой панели модуля *Nonparametrics* выбираем *Comparing multiple indep. samples (groups).*

В появившемся окне (рис. 7.14) выбираем *Variables* и задаём переменные (рис. 7.15); затем нажимаем OK.

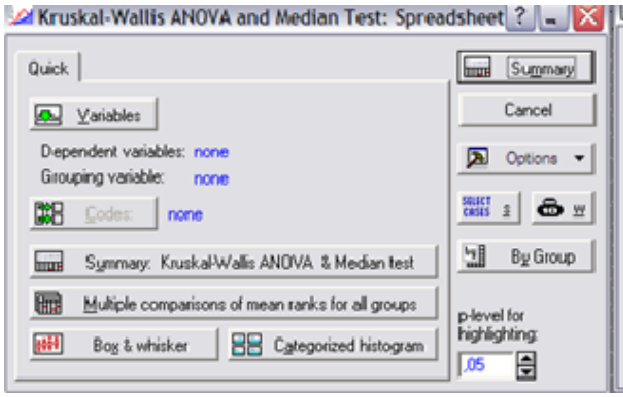


Рис. 7.14 – Окно *Kruskal-Wallis Anova and Median test*

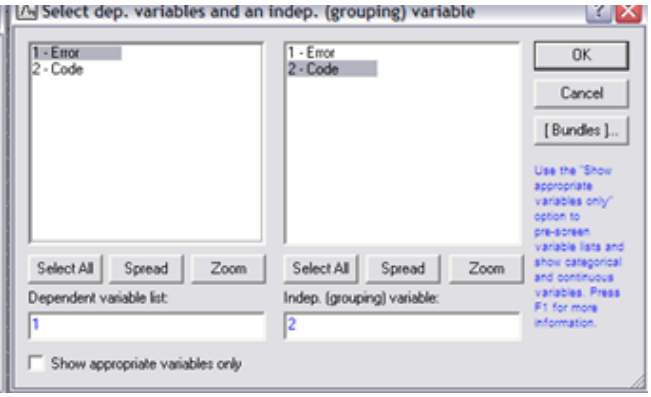


Рис. 7.15 – Окно выбора переменных

Далее нажимаем *Codes* и выбираем коды для группируемых переменных, щёлкнув по кнопке *All* (рис. 7.16):

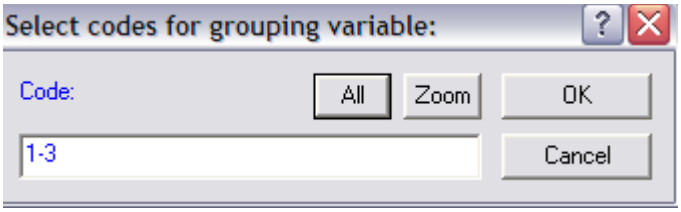
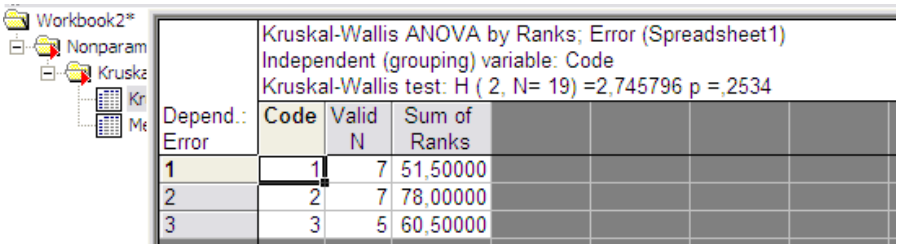


Рис. 7.16 – Окно выбора кода

Нажимаем *Summary* и получаем следующую таблицу результатов (рис. 7.17):



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. 7.17 – 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 7.4. 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 7.1. Наибольший балл соответствует плёнке самого лучшего качества.

Таблица 7.1

Вид плёнки	**Эксперты**		
	1	**2**	**3**
П1	2	3	2
П2	5	4	5
П3	3	3	3
П4	4	5	5

Требуется определить, различаются ли виды плёнок и согласованы ли оценки экспертов. Задание выполнить в пакете *Statistica*.

Решение. Введём исходные данные (рис. 7.18):

	1 P1	2 P2	3 P3	4 P4
1	2	5	3	4
2	3	4	3	5
3	2	5	3	5

Рис. 7.18 – 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. 7.19):

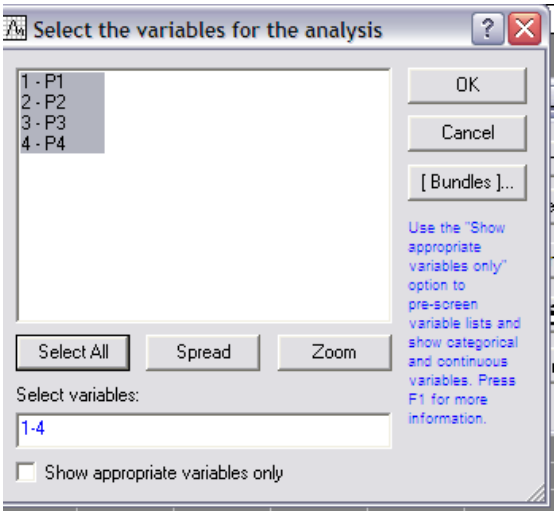


Рис. 7.19 – Окно выбора переменных

В появившемся окне *Friedman ANOVA by ranks* (Двухфакторный анализ Фридмана) нажимаем *Summary* и получаем следующую таблицу результатов (рис. 7.20):

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. 7.20. – 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.

7.3. Задания для самостоятельной работы.

Задание 1.

Вариант 1. Используя критерий знаков и знако-ранговый критерий Вилкоксона проверить гипотезу H_0: $F_1\ x\ = F_2\ x\ $ об однородности двух выборок (в первой строке приведены варианты первой выборки, во второй строке – варианты второй выборки). Принять $\alpha = 0{,}01$.

X	75	65	87	80	84	96	63	58	66	65	88	90
Y	80	70	88	75	90	93	65	60	66	70	82	91

Задание выполнить в пакете STATISTICA.

Вариант 2. Ниже приводится время (в секундах) решения контрольных задач одиннадцатью учащимися до и после специальных упражнений по устному счёту. Можно ли считать, что эти упражнения улучшили способности учащихся в решении задач. Принять $\alpha = 0{,}05$.

До упражнений	87	61	98	90	93	74	83	72	81	75	83
После упражнений	50	45	79	90	88	65	52	79	84	61	52

Задание выполнить в пакете STATISTICA (используя критерий знаков и знако-ранговый критерий Вилкоксона).

Вариант 3. Для определения качества технологической операции регулярно осуществляются проверки, которые состоят в измерении одного параметра изделия, прошедшего данную операцию. Имеются данные за 2 дня:

X	82	74	64	72	84	68	76	88	70	60
Y	52	63	72	64	48	70	79	68	70	54

Необходимо оценить стабильность контролируемой операции (проверить однородность выборок) при уровне значимости $\alpha = 0{,}05$. Задание выполнить в пакете STATISTICA (используя критерий знаков и знако-ранговый критерий Вилкоксона).

Вариант 4. Используя критерий знаков и знако-ранговый критерий Вилкоксона проверить гипотезу H_0: $F_1\ x\ = F_2\ x\ $ об однородности двух выборок (в первой строке приведены варианты первой выборки, во второй строке – варианты второй выборки). Принять $\alpha = 0{,}01$.

X	64	70	54	55	71	51	53	83	72	59	51
Y	62	61	53	50	68	51	49	80	75	65	50

Задание выполнить в пакете STATISTICA.

Вариант 5. Используя критерий знаков и знако-ранговый критерий Вилкоксона проверить гипотезу H_0: $F_1\ x\ = F_2\ x\ $ об однородности двух выборок (в первой строке приведены варианты первой выборки, во второй строке – варианты второй выборки). Принять $\alpha = 0{,}01$.

X	59	60	45	78	64	82	68	60	62	79	61
Y	62	65	44	74	60	83	64	66	62	82	66

Задание выполнить в пакете STATISTICA

Вариант 6. Данные о производительности выпуска стиральных машин на двух предприятиях представлены в таблице:

X	80	74	73	71	89	76	90	76	60
Y	60	63	72	69	49	74	80	76	55

Можно ли распределение производительности на обоих предприятиях считать различным при $\alpha = 0.05$. Задание выполнить в пакете STATISTICA (используя критерий знаков и знако-ранговый критерий Вилкоксона).

Вариант 7. Результаты измерения уровня тревожности до и после проведения тренинга в группе испытуемых отображены в таблице

№ испытуемого	1	2	3	4	5	6	7	8	9	10	11	12	13	14
Уровень тревожности до тренинга	30	39	34	35	40	35	22	22	31	23	16	34	33	34
Уровень тревожности после тренинга	34	39	26	33	34	40	25	23	32	24	15	27	35	37

Определить, является ли изменение уровня тревожности статистически значимым. Принять $\alpha = 0{,}05$. Задание выполнить в пакете STATISTICA (используя критерий знаков и знако-ранговый критерий Вилкоксона).

Вариант 8. Используя критерий знаков и знако-ранговый критерий Вилкоксона проверить гипотезу H_0: $F_1\ x\ = F_2\ x\ $ об однородности двух выборок (в первой строке приведены варианты первой выборки, во второй строке – варианты второй выборки). Принять $\alpha = 0{,}01$

X	**53**	**83**	**50**	**65**	**58**	**50**	**58**	**79**	**50**	**84**	**70**
Y	**57**	**75**	**55**	**69**	**66**	**72**	**60**	**80**	**51**	**85**	**68**

Задание выполнить в пакете STATISTICA.

Задание 2. Используя критерии Манна-Уитни, Вальда-Вольфовица и Колмогорова-Смирнова, проверить (в пакете STATISTICA) 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.

Вариант 1 $\alpha = 0{,}05$													
x_i	0,2	0,3	0,5	0,8	1	1,3							
y_i	0,1	0,4	0,6	0,7	0,9	1,4	1,7	1,8	1,9				
Вариант 2 $\alpha = 0{,}01$													
x_i	3	4	6	10	13	17	19	20					
y_i	1	2	5	7	16	20	22	19	15				
Вариант 3 $\alpha = 0{,}1$													
x_i	28	33	39	40	41	42	45	46	47				
y_i	34	40	41	42	43	44	46	48	49	92			
Вариант 4 $\alpha = 0{,}01$													
x_i	560	580	600	420	530	490	580	470					
y_i	692	700	621	640	561	680	630						
Вариант 5 $\alpha = 0{,}05$													
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 6 $\alpha = 0{,}05$													
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	0,60	0,66	0,73	0,87	0,94			
Вариант 7 $\alpha = 0{,}05$													
x_i	95,6	94,9	96,2	95,1	95,8	96,3							
y_i	93,3	92,1	94,7	90,1	95,6	95,4	90	94,7					
Вариант 8 $\alpha = 0{,}05$													
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				

Задание 3.

m групп водителей обучались по различным методикам. После окончания срока обучения был проведён тестовый контроль над случайно отобранными водителями из каждой группы. Получены следующие результаты (задать самостоятельно):

№ группы	**Число ошибок, допущенных водителями**	**Сумма ошибок по каждой группе**	**Число контролируемых водителей**
…	…	…	…
m			

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.

Задание 4. Киноплёнка n видов была представлена m экспертам для определения лучшей из них. Каждому эксперту предложили упорядочить плёнки по степени предпочтения. Баллы (ранги), поставленные экспертами, приведены в таблице (задать самостоятельно). Наибольший балл соответствует плёнке самого лучшего качества.

Вид плёнки	**Эксперты**		
		…	$\boldsymbol{n}$
…			
m			

Требуется определить, различаются ли виды плёнок и согласованы ли оценки экспертов (вычислить статистику Фридмана и коэффициент конкордации W по данным задачи). Задание выполнить в пакете STATISTICA.

Лабораторная работа № 8.

АНАЛИЗ И ПРОГНОЗИРОВАНИЕ ВРЕМЕННЫХ РЯДОВ. МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

Цель работы: Освоить методологию анализа, моделирования и прогнозирования стационарных временных рядов. Создание модели множественной регрессии для прогноза случайной величины. Проверка прогноза на зависимом и независимом материале. Оценка качества модели.

Используемые программные средства: MS Excel 2010 (2016), STATISTICA 8.0.

8.1. Краткие теоретические сведения.

*Основные понятия и определения***.** Рассмотрим случайный объект или явление, характеристика Y которого меняется во времени. Выполнив n наблюдений над этим объектом в равноотстоящие моменты времени $t_1, t_2, \dots, t_n$, получим упорядоченную последовательность чисел:

$$y\ t_1\ , y\ t_2\ , \dots , y\ t_n$$

которая называется *временным (динамическим) рядом* или *случайной последовательностью.* Чаще используется более компактная форма временных рядов $y_1, y_2, \dots, y_n$ (здесь $y_i \equiv y\ t_i\ , i = 1, n$). 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 $x_1, x_2, \dots, x_n$ 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 $y_1, y_2, \dots, y_{i-1}$, зафиксированных до этого момента.
2. Элементы $x_1, x_2, \dots, x_n$ 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 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:

$$y_t = \frac{y_{t-m+1} + y_{t-m+2} + \cdots + y_t}{m},$$

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:

$$Q_1 = y_1;\ Q_t = \alpha \cdot y_t + \ 1 - \alpha \ \cdot Q_{t-1} \quad t = 2, n\ ,$$

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 $0 \leq \alpha \leq 1$.

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 y_0. В качестве начального сглаженного значения обычно используют первый член исследуемого временного ряда, т.е. $y_0 = y_1$.

В пакете *Excel* для сглаживания временных рядов используются: процедуры *Скользящее среднее* и *Экспоненциальное сглаживание,* входящие в *Пакет анализа*. Диалоговые окна этих процедур приведены на рис. 8.1.

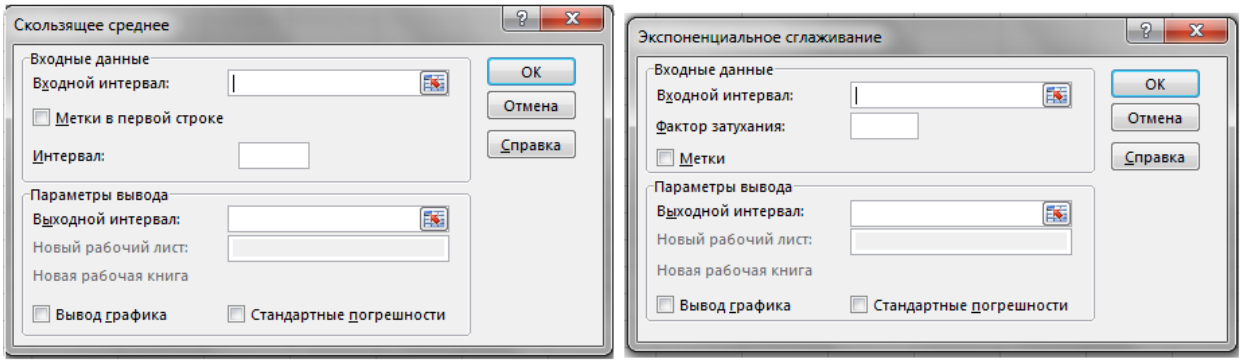


Рис. 8.1 – Диалоговые окна процедур *Скользящее среднее* и *Экспоненциальное сглаживание*

Диалоговые окна рассматриваемых процедур содержат следующие специфические для этих процедур элементы управления:

- поле ввода *Интервал* диалогового окна *Скользящее среднее*. В него вводится размер m окна сглаживания (по умолчанию $m = 3$);

- поле ввода *Фактор затухания* диалогового окна *Экспоненциальное сглаживание*. В этом поле вводится фактор затухания $\beta = 1 - \alpha$;
- флажок *Вывод графика*. При установке этого флажка на экран выводятся графики исходного и сглаженного временных рядов;
- флажок *Стандартные погрешности*. Устанавливается при необходимости получения стандартных погрешностей сглаживания.

Входной интервал, содержащий элементы исследуемого временного ряда, должен состоять из одного столбца, «высота» n которого равна числу элементов этого ряда.

Выходной интервал состоит по крайней мере из одного столбца, содержащего элементы сглаженного ряда. Высота этого столбца равна высоте входного интервала. При установке флажка *Стандартные погрешности* в выходном интервале появляется еще один столбец – столбец стандартных погрешностей. В точках, для которых нельзя вычислить сглаженные значения и стандартные погрешности, процедура выводит сообщения #Н/Д! – нет данных.

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 существенная 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* $y\ t\ = a_0 + a_1 t$. Эта модель описывает тренд, скорость изменения которого постоянна и равна a_1. При $a_1 > 0$ тренд равномерно возрастает, при $a_1 < 0$ – равномерно убывает.
- *Логарифмическая* $y\ t\ = a_0 + a_1 \ln t$, описывающая тренд с постепенным уменьшением скорости роста.
- *Полиномиальная* $y\ t\ = a_0 + a_1 t + a_2 t^2 + \cdots + a_m t^m$, где m – степень полинома. Частный случай этой модели – полином второй степени $y\ t\ = a_0 + a_1 t + a_2 t^2$ описывает тренд с постоянным ускорением из-

менения, равным $2a_2$. При $a_2 > 0$ скорость изменения тренда возрастает, при $a_2 < 0$ – убывает. Полином третьей степени $y\ t\ = a_0 + a_1 t + a_2 t^2 + a_3 t^3$ описывает тренд с переменным изменением ускорения. При $a_3 > 0$ ускорение возрастает, при $a_3 < 0$ – убывает.

- *Степенная* $y\ t\ = a_0 t^{a_1}$.
- *Экспоненциальная* $y\ t\ = a_0 e^{a_1 t} = a_0 exp\ a_1 t$, 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 $\sum_{i=1}^{n} y_i - y\ t_i\ ^2$, которая была минимизирована при расчете параметров тренда.

Но чаще точность оценивается на основании *коэффициента детерминации* R^2:

$$R^2 = 1 - \frac{\sum_{i=1}^{n} y_i - y\ t_i\ ^2}{\sum_{i=1}^{n} y_i - y\ ^2},$$

где n – количество уровней временного ряда (число наблюдений); y_i – фактическое значение показателя в момент времени t_i; $y\ t_i\ $ – теоретическое (рассчитанное по тренду) значение показателя в момент времени t_i; y – среднее арифметическое фактических значений, которое рассчитывается по формуле $y = \frac{1}{n}\ _i y_i$.

Коэффициент детерминации всегда удовлетворяет условию $0 \leq R^2 \leq 1$. 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.

Эффективным и очень удобным в использовании средством аналитиче-

ского сглаживания является функция *Добавить линию тренда*, входящую в комплекс графических средств табличного процессора *Excel*.

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

$$Y = a_0 + a_1X_1 + a_2X_2 + \cdots + a_mX_m$$

где Y – *зависимая переменная* (предиктант, отклик); a_0 – *константа*; X_i – независимые переменные (предикторы, факторы); $i = 1,2, \dots, m$; m – количество предикторов.

Слово «предиктор» произошло от англ. *predict* – предсказывать. Процедуры множественной регрессии будут оценивать (вычислять) *параметры уравнения*, то есть коэффициенты $a_0, a_1, \dots, a_m$. Величины $a_1, a_2 \ \dots, a_m$ 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 независимыми переменными корреляционная матрица становится вырождающейся, её определитель стремится к нулю, и возникают трудности в вычислении коэффициентов уравнения регрессии. В этом случае надо исключать дублирующие независимые переменными.

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 фактора 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 прямой. 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 статистике принято считать, что уравнение регрессии с данным набором факторов можно использовать, если факторы 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 изме-

чивость соответствующей переменной.

Интерпретация коэффициента множественной корреляции R. Обычно, степень зависимости двух или более факторов 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 факторов.

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.

8.2. Практическая часть.

Контрольный пример 8.1. Ниже приведены данные о числе сообщений, переданных в течение 42 суток (6 недель) в одной из радиосетей морской связи.

Таблица 8.1

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.

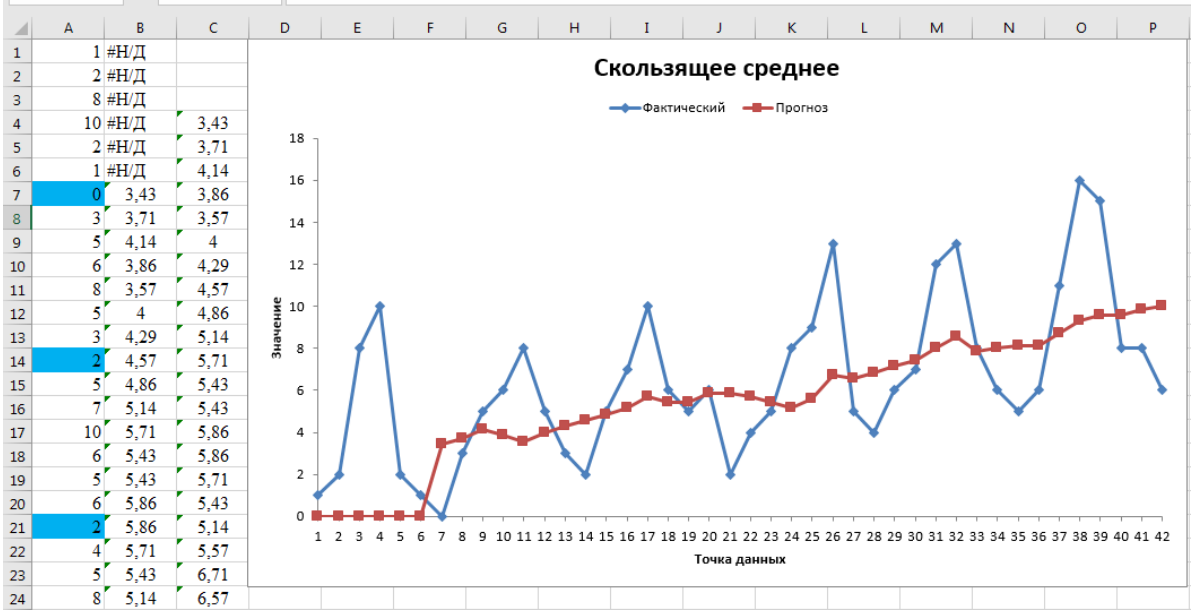
В диапазон A1: A42 листа Excel ведем значения временного ряда из при-

веденной таблицы (на рис. 9.2 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. 8.2 – Решение примера 8.1 (сглаживание с помощью скользящего среднего)

Справа от столбца с исходными данными в диапазоне B1:B42, появятся столбец адаптивных скользящих средних y_i и график исследуемого временного ряда с наложенным на него графиком адаптивных скользящих средних $y_i; i = 7{,}42$ (см. рис. 8.2).

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. 8.3. На этом же рисунке приведен график адаптивного скользящего среднего.

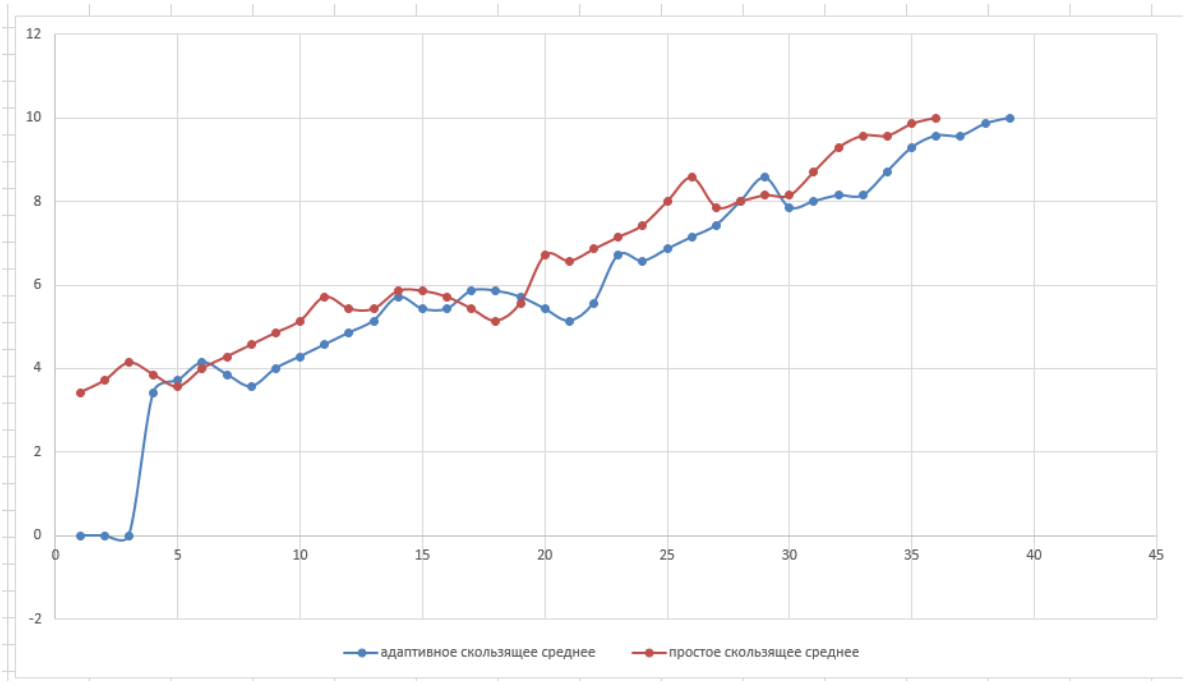


Рис. 8.3. – Сравнение простого y_i и адаптивного y_i скользящих средних

На рис. 8.4 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. 8.4 – Решение контрольного примера 8.1 (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. 8.5 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»).

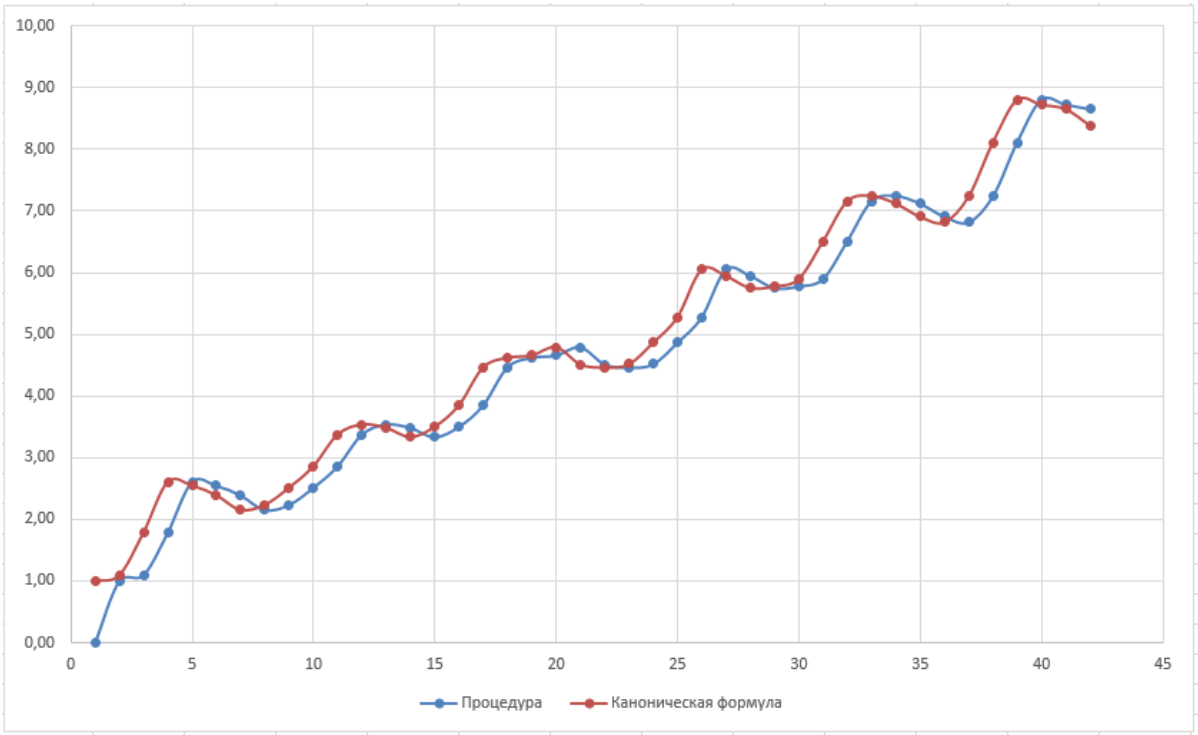


Рис. 8.5 – Сравнение «канонического» экспоненциального сглаживания и сглаживания, выполненного с помощью процедуры *Экспоненциальное сглаживание*

Контрольный пример 8.2. Имеются данные об объемах продаж некоторой фирмы (табл.8.2). С помощью графика подобрать линию тренда, которая лучше всего описывает фактические данные и на ее основе сделать прогноз на три недели вперед.

Таблица 8.2

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. 8.6)

	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. 8.6 – 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. 8.7). Уравнение и коэффициент детерминации можно выделить щелчком левой кнопки мыши и перетащить ее на то место графика, где их лучше видно.

Рис. 8.7 – 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 8.3 занесем линию тренда и соответствующее значение R^2.

Таблица 8.3

№	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	$y = 14{,}483\ x^{0.5858}$	$R^2 = 0{,}9129$
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. 8.8).

Рис. 8.8 – 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 A. Столбец C озаглавим «Теоретические значения» и занесем в него формулы расчета по выбранной функции тренда. В ячейку C2 запишем формулу $= 16{,}257 * \text{EXP}\ 0{,}1361 * \text{A2}$. Эта формула копируется методом автозаполнения в ячейки C3:C15, т.е. теоретические значе-

ния рассчитываются для всех моментов времени в прошлом и прогнозируемом будущем (рис. 8.9).

В результате получим в ячейках C13:C15 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. 8.9 – Прогнозирование продаж на три недели вперед

Контрольный пример 8.3. 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 8.4 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 8.4

№	1	2	3	4	5	6	7	8	9	10	11	12	13	14
X_1	1	1	1.1	1.1	1.1	1.1	1	1	1.2	1.2	0.6	0.6	0.7	0.7
X_2	1.3	1.3	1.4	1.4	1.5	1.5	1.4	1.5	1.6	1.7	1	1	1.1	1.15
Y	1160	1155	1158	1157	1160	1161	1157	1159	1256	1260	1040	1039	1039	1040
№	15	16	17	18	19	20	21	22	23	24	25	26	27	28
X_1	0.75	0.7	0.7	0.7	0.8	0.8	0.78	0.8	0.78	0.78	0.8	0.8	0.75	0.78
X_2	1.2	1.2	1.3	1.3	1.4	1.4	1.5	1.5	1.5	1.6	1.7	1.8	1.8	1.9

Y	1040	1039	1040	1039	1140	1138	1240	1239	1241	1240	1239	1239	1240	1238

Решение. Введем исходные данные, расположив каждую случайную величину в отдельном столбце (на рис. 8.10 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. 8.10 – 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 $= \text{КОРРЕЛ}(\text{A2:A29};\text{C2:C29})$ и $= \text{КОРРЕЛ}(\text{B2:B29};\text{C2:C29})$. 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 $y = a_0 + a_1X + a_2X_2$. В меню *Данные* выберем *Анализ данных – Регрессия*. Заполним диалоговое окно, как показано ни рис. 8.11.

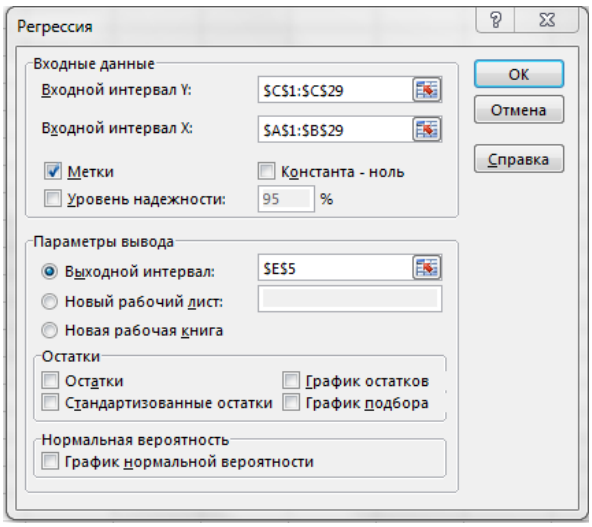


Рис. 8.11 – Пример заполнения диалогового окна *Регрессия*

Результаты работы процедуры *Регрессия* из Пакета Анализа показаны на рис. 8.12

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. 8.12 – 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:

$$y = 672{,}5 + 79{,}7 \cdot X + 288{,}9 \cdot X_2$$

Для каждого коэффициента рассчитана также стандартная ошибка и выборочное значение -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 $= \text{СТЬЮДЕНТ.ОБР.2Х}(0{,}05; 26)$ (здесь $26 = n - 2$):

H	I	J
	2,056	

Сравнивая это значение с -статистикой для каждого параметра, убеждаемся, что для a_0 и a_2 выполняется условие $t > t_{\text{кр}}$ ($13{,}118 > 2{,}056$ и $8{,}241 > 2{,}056$), 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. 8.13. 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. 8.13 – Расчет прогноза заболеваемости

Решим данную задачу с применением пакета *Statistica*. Образуем таблицу с 3 столбцами и 28 строками. Вводим в таблицу исходные данные (рис.8.14):

	1 X1	2 X2	3 Y				
1	1	1,3	1160	15	0,75	1,2	1040
2	1	1,3	1155	16	0,7	1,2	1039
3	1,1	1,4	1158	17	0,7	1,3	1040
4	1,1	1,4	1157	18	0,7	1,3	1039
5	1,1	1,5	1160	19	0,8	1,4	1140
6	1,1	1,5	1161	20	0,8	1,4	1138
7	1	1,4	1157	21	0,78	1,5	1240
8	1	1,5	1159	22	0,8	1,5	1239
9	1,2	1,6	1256	23	0,78	1,5	1241
10	1,2	1,7	1260	24	0,78	1,6	1240
11	0,6	1	1040	25	0,8	1,7	1239
12	0,6	1	1039	26	0,8	1,8	1239
13	0,7	1,1	1039	27	0,75	1,8	1240
14	0,7	1,15	1040	28	0,78	1,9	1238

Рис. 8.14 – 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. 8.15).

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. 8.15).

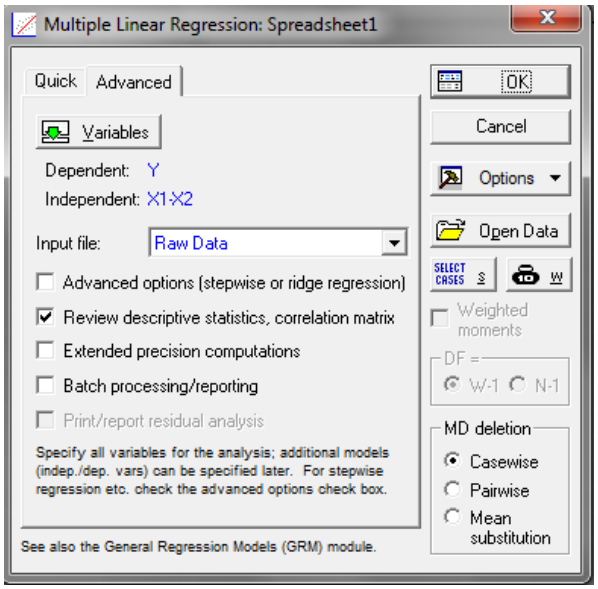


Рис. 8.15 – Стартовая панель модуля множественной регрессии

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. 8.16 – 1), а кнопкой *Корреляция* (*Correlation*) получить матрицу коэффициентов корреляции (рис. 8.16 – 2).

Variable	Means and Standard Deviations (Spreadsheet1)		
	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. 8.16 – 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. 8.15).

В окне *Review Descriptive Statistics* нажмем кнопку ОК.

Диалоговое окно результатов (*Multiple Regression Results)* состоит из двух частей – информационной и функциональной (рис. 8.17).

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, соответствующему F-статистике, можно сделать предварительное заключение о пригодности уравнения. Если уровень значимости $p \leq p\text{крит}$ (по умолчанию $p\text{крит} = 0{,}05$), то уравнение регрессии с данным набором предикторов пользоваться не имеет смысла.

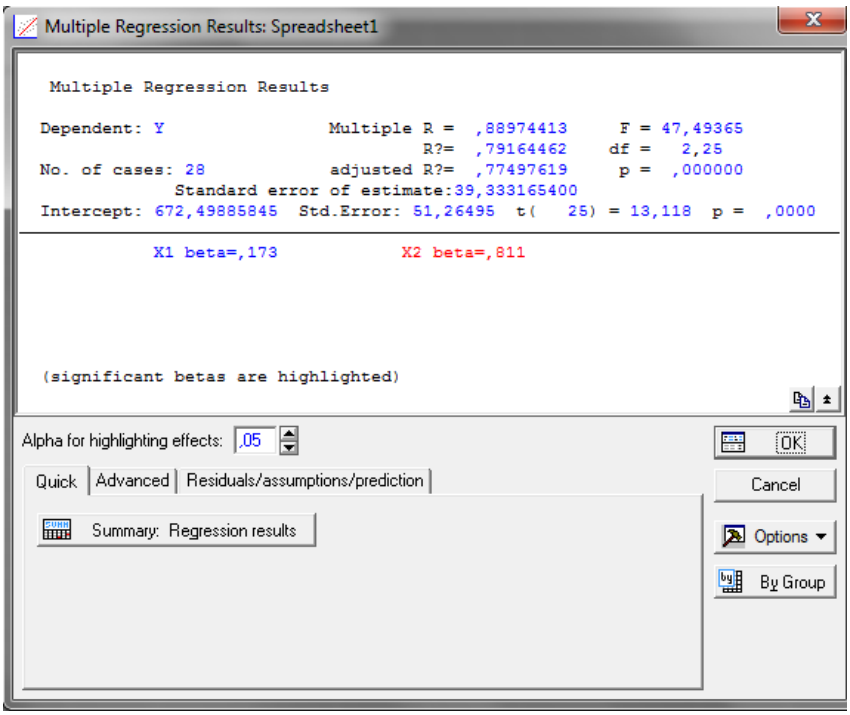


Рис. 8.17 – Окно результатов

Важную информацию несут bet-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 8.17, фактор X2 вносит в прогноз величины Y значительно больший вклад, чем X1. Значимые bet-коэффициенты выделяются красным цветом. Факторы, имеющие незначимые bet-коэффициенты, из уравнения регрессии удаляются как неинформативные.

Нажав на кнопку *Summary: Regression results*, получим таблицу результатов (рис.8.18):

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. 8.18 Результаты регрессионного анализа

Обобщённый коэффициент корреляции равен: $R = 0{,}8897$. Остаточная дисперсия – $D_{ост} = 39{,}33^2 = 1546{,}8$. В столбце B указаны оценки неизвестных коэффициентов. Таким образом, оценка неизвестной функции регрессии имеет вид $y = 672{,}5 + 79{,}7x_1 + 288{,}9x_2$.

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 $DW1$ и верхним $DW2$.

- Если $d < DW1$ (или $4 - d < DW1$), то в остатках имеется автокорреляция на заданном уровне значимости.
- Если $d > DW2$ (или $4 - d > DW2$), то автокорреляция отсутствует.
- Если $DW1 < d < DW2$, то случай сомнительный, нужны дополнительные исследования.

Когда расчётное значение статистики d превышает 2, то с $DW1$ и $DW2$ 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 p. В таблице 8.5 приводится таблица критических точек статистики Дарбина-Уотсона для уровня значимости $\alpha = 0{,}05$:

Вычислим статистику Дарбина-Уотсона для остатков в диалоговом окне *Анализ остатков* (*Residual Analysis*) на вкладке *Дополнительно* (*Advanced*) (рис. 8.19, 8.20).

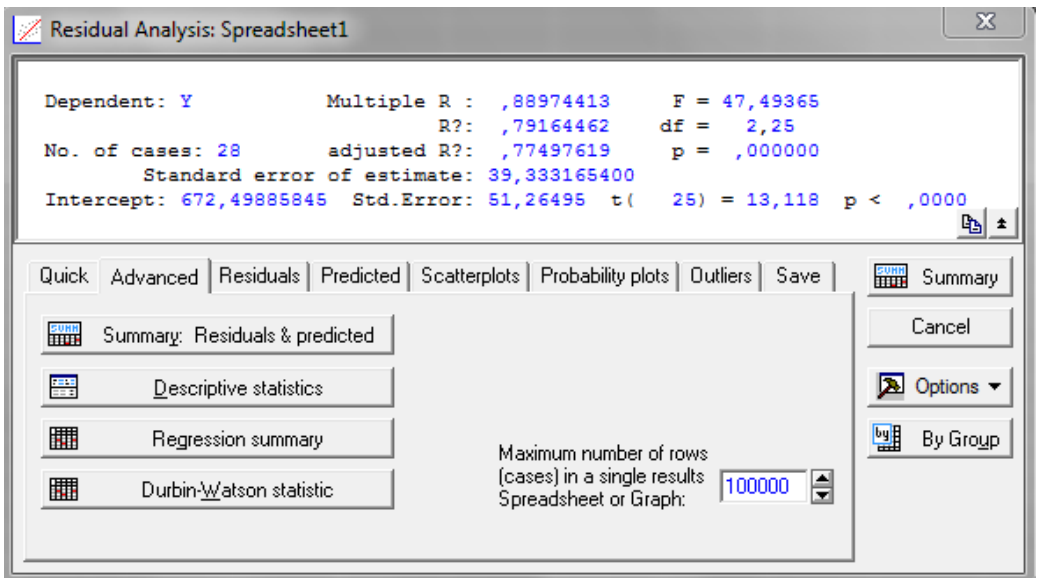


Рис. 8.19 – Окно анализа остатков

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. 8.20 – Окно с вычисленной статистикой Дарбина -Уотсона

Так как $\mathrm{DW1} = 1{,}25;\ \mathrm{DW2} = 1{,}56$ и $d < \mathrm{DW1}$, то гипотеза о независимости случайных отклонений отвергается (следовательно, присутствует положительная автокорреляция.

Для прогноза в окне результатов (рис. 8.17) на вкладке *Остатки/ предсказанные/ наблюденные значения* ($Residuals\ /\ assumptions\ /\ prediction$) имеется кнопка с зеленой стрелкой *Предсказать зависимую переменную* ($Predict\ dependent\ variable$). Она вызывает специальное окно для прогноза (рис. 8.21), в котором надо задать числовые значения факторов X_i.

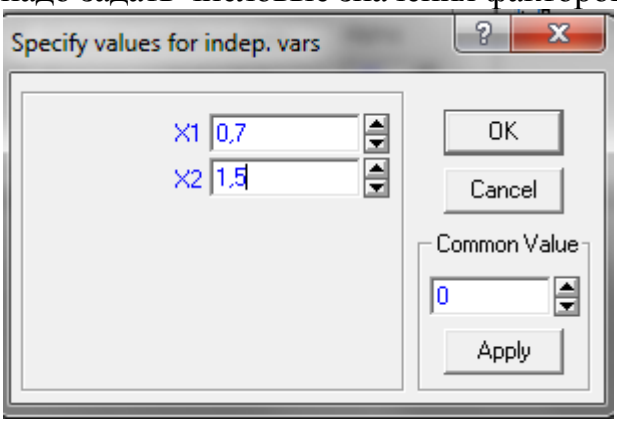


Рис. 8.21 – Окно задания прогноза

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. 8.22).

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. 8.22 – Окно с результатами прогноза по уравнению регрессии

По информации из этой таблицы можно выписать уравнение регрессии $y = 79{,}7X_1 + 288{,}9X_2 + 672{,}5$.

Таблица 8.5

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
1	2	3	4	5	6	7	8	9	10	11
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	1,25	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
31	1,36	1,5	1,3	1,57	1,23	1,65	1,16	1,74	1,09	1,83
32	1,37	1,5	1,31	1,57	1,24	1,65	1,18	1,73	1,11	1,82
33	1,38	1,51	1,32	1,58	1,25	1,65	1,19	1,73	1,13	1,81
34	1,39	1,51	1,33	1,58	1,27	1,65	1,21	1,73	1,05	1,81
35	1,4	1,52	1,34	1,58	1,28	1,65	1,22	1,73	1,16	1,8
36	1,41	1,52	1,35	1,59	1,29	1,65	1,24	1,73	1,18	1,8
37	1,42	1,53	1,36	1,59	1,31	1,66	1,25	1,72	1,19	1,8
38	1,43	1,54	1,37	1,59	1,32	1,66	1,26	1,72	1,21	1,78
39	1,43	1,54	1,38	1,6	1,33	1,66	1,27	1,72	1,22	1,79
40	1,44	1,54	1,39	1,6	1,34	1,66	1,29	1,72	1,23	1,79
45	1,46	1,57	1,43	1,62	1,38	1,67	1,34	1,72	1,29	1,79
50	1,5	1,58	1,46	1,63	1,42	1,67	1,36	1,72	1,34	1,77
55	1,51	1,6	1,49	1,64	1,45	1,68	1,41	1,72	1,38	1,77

60	1,55	1,62	1,51	1,65	1,48	1,69	1,44	1,73	1,41	1,77
65	1,57	1,63	1,54	1,66	1,5	1,7	1,47	1,73	1,44	1,77

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	171	147	169	162	186	181	168	222	195	
Вариант 2 $m = 5;\ \alpha = 0{,}4$										
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 3 $m = 3;\ \alpha = 0{,}3$										
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 4 $m = 3;\ \alpha = 0{,}4$										
t	1	2	3	4	5	6	7	8	9	10
x	4633	4742	4980	5187	5107	5358	5299	5228	5345	5244
Вариант 5 $m = 5;\ \alpha = 0{,}2$										
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 6 $m = 3;\ \alpha = 0{,}3$										
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 7 $m = 4;\ \alpha = 0{,}4$										
t	1	2	3	4	5	6	7	8	9	10
x	67	59	47	62	64	46	51	37	63	56

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	10
Количество продаж	2	1	4	4	6	8	7	9	12	11
Вариант 2										
Неделя	1	2	3	4	5	6	7	8	9	10
Количество продаж	15	15	16	22	32	39	49	58	65	67
Вариант 3										
Неделя	1	2	3	4	5	6	7			
Количество продаж	659	656.7	645.1	633.6	621.7	605.6	585.03			
Вариант 4										
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 5										
Неделя	1	2	3	4	5	6	7	8	9	
Количество продаж	3	10	11	18	21	30	32	37	42	
Вариант 6										
Неделя	1	2	3	4	5	6	7	8	9	10
Количество продаж	40	42	43	46	49	50	56	62	65	63
Вариант 7										
Неделя	1	2	3	4	5	6	7	8	9	10
Количество продаж	18	25	27	26	30	37	35	40	48	53

Задача 3. По представленным ниже данным построить уравнение множественной линейной регрессии и оценить его качество. Задание выполнить в пакетах STATISTICA и *Excel*

Вариант 1										
X_1	1	4	0	5	-3	3	-5	-1	2	-2
X_2	4	-6	2	-4	12	-2	14	6	0	8
Y	-4	-5	4	-1	4	0	5	1	2	7
Вариант 2										
X_1	5,84	3,82	6,19	9,22	7,87	6,29	4,43	8,91	5,34	2,21
X_2	6,04	6,33	4,86	5,91	4,96	5,58	6,15	6,13	4,65	5,49
Y	79,31	57,43	60,66	92,6	90,1	71,3	70,5	91,5	68,31	58,56
Вариант 3										
X_1	0	44	4	61	35	64	13	56	18	2
X_2	14	0	29	34	54	16	44	59	49	32
Y	0,5	47,2	8	63,8	18,2	47,5	0	60,9	19,2	9
Вариант 4										
X_1	3,05	2,2	0,65	1,65	1,92	1,92	0,89	0,75	2,79	0,44
X_2	7,92	4,71	8,09	8,35	6,24	4,39	6,95	3,67	2,88	3,71
Y	5,64	2,93	5,16	5,62	4,05	2,61	4,33	1,75	1,65	1,70
Вариант 5										
X_1	5,14	5,59	4,33	4,59	4,21	3,78	4,23	5,61	4,87	3,87
X_2	4,23	1,4	4,07	2,93	3,44	1,09	1,82	2,43	3,85	0,97
Y	65,72	58,05	60,05	55,8	50,8	47,69	44,49	59,7	56,81	45,82
Вариант 6										
X_1	0,12	3,48	4,45	6,19	1,81	3,81	0,84	2,08	1,28	5,44
X_2	2,91	2,94	6,35	6,58	3,8	6,43	0,57	5,96	3,4	4,55
Y	82,16	61,02	44,56	82,5	99,2	70,24	63,23	66,5	48,35	40,24
Вариант 7										
X_1	0,12	3,48	4,45	6,19	1,81	3,81	0,84	2,08	1,28	5,44
X_2	6,04	6,33	4,86	5,91	4,96	5,58	6,15	6,13	4,65	5,49
Y	82,2	61,02	44,56	82,5	99,17	70,2	63,2	66,5	48,4	40,2

Вопросы к зачету по курсу «Прикладная математика»

1. Генеральная совокупность, выборка, репрезентативность, полигон, гистограмма.
2. Точечные оценки параметров распределения и их свойства: несмещенность, состоятельность, эффективность. Оценки математического ожидания и дисперсии
3. Интервальные оценки параметров, надежность оценки.
4. Построение доверительных интервалов для математического ожидания и дисперсии нормального распределения.
5. Проверка статистических гипотез. Нулевая, конкурирующая, простая и сложная гипотезы. Ошибки первого и второго рода.
6. Критерии проверки гипотез. Уровень значимости и критические области. Схема проверки статистических гипотез.
7. Критерий χ^2. Проверка гипотез о характере распределения случайной величины (равномерного, показательного, биномиального и других распределений).
8. Приближенная проверка на нормальность (графическая, с помощью коэффициентов асимметрии и эксцесса, с использованием σ_B).
9. Критерий χ^2. Проверка гипотезы о нормальном распределении случайной величины. Правило Романовского.
10. Критерий Колмогорова.
11. Проверка гипотезы о значении математического ожидания нормально распределённой случайной величины.
12. Проверка гипотез равенства математических ожиданий двух случайных величин, распределённых нормально.
13. Проверка гипотезы о дисперсиях двух случайных величин, распределённых по нормальному закону. Критерий Фишера.
14. Проверка гипотез о дисперсиях нескольких случайных величин, имеющих нормальное распределение. Критерии Бартлетта и Кохрена.
15. Непараметрическое сравнение выборочных статистик. Критерий Манна-Уитни и двухвыборочный критерий Вилкоксона.
16. Непараметрические методы математической статистики. Проверка однородности двух выборок – критерий знаков и знако-ранговый критерий Вилкоксона.
17. Функциональная, статистическая и корреляционная зависимости.
18. Линейная регрессия. Метод наименьших квадратов.
19. Остаточная дисперсия. Коэффициент детерминации. Адекватность линейной регрессии результатам наблюдений.
20. Ковариация и выборочный коэффициент корреляции. Проверка статистических гипотез о корреляционной зависимости.
21. Выборочный коэффициент ранговой корреляции Спирмена, его свойства. Проверка гипотезы о его значимости.

22. Выборочный коэффициент ранговой корреляции Кендалла, его свойства. Проверка гипотезы о его значимости.
23. Выборочное корреляционное отношение и его свойства.
24. Нелинейная регрессия.
25. Некоторые нелинейные задачи, сводящиеся к линейным моделям.
26. Множественная регрессия. Построение линейной регрессионной модели. Коэффициент множественной корреляции, его свойства.
27. Автокорреляция остатков. Критерий Дарбина-Уотсона.
28. Дисперсионный анализ. Основные понятия.
29. Однофакторный дисперсионный анализ.
30. Двухфакторный дисперсионный анализ Основные понятия.
31. Однофакторный непараметрический анализ. Критерий Краскелла-Уоллиса.
32. Двухфакторный непараметрический анализ. Критерий Фридмана. Коэффициент конкордации (согласованности).
33. Временные ряды. Основные понятия и определения.
34. Стационарные временные ряды и их характеристики. Автокорреляционная функция.
35. Аналитическое выравнивание (сглаживание) временного ряда.
36. Прогнозирование на основе моделей временных рядов.
37. Понятие об авторегрессионных моделях и моделях скользящей средней.
38. Временной ряд, тренд, трендовая модель. Получение трендовой модели средствами Excel.

Список литературы

1. 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.
2. Савич, Л.К. Теория вероятностей и математическая статистика: учебн. Пособие для студентов экон. специальностей учреждений, обеспечивающих получение высшего образования / Л.К. Савич, Н.А. Смольская; науч. ред. О.И. Лаврова. – Мн.: Адукацыя і выхаванне, 2006. – 208 с.: ил.
3. Лисьев, В.П. Теория вероятностей и математическая статистика: Учебное пособие/ Московский государственный университет экономики, статистики и информатики. – М., 2006. – 199 с.
4. Иванов, О.В. Статистика / Учебный курс для социологов и менеджеров. Часть 1. Описательная статистика. Теоретико-вероятностные основания статистического вывода. – М. 2005. – 187 с.
5. Иванов, О.В. Статистика / Учебный курс для социологов и менеджеров. Часть 2. Доверительные интервалы. Проверка гипотез. Методы и их применение. – М. 2005. – 220 с.
6. Вадзинский, Ратмир. Статистические вычисления в среде Excel. Библиотека пользователя. – СПб.: Питер, 2008. – 608 с.: ил. – (Серия «Библиотека пользователя»).
7. Евдокимова, Г.С. Математическая статистика в примерах и задачах: учебное пособие / Г.С. Евдокимова; Смол. гос. ун-т. – Смоленск: Изд-во СмолГУ, 2014. – 98 с.
8. Жученко, Ю.М. Математическая статистика в биологии и химии: учебное пособие для студентов вузов по специальности 1-31 01 01 «Биология» / Ю.М. Жученко; М-во образования РБ, Гомельский гос. университет им. Ф. Скорины. – Гомель: ГГУ им. Ф. Скорины, 2010. – 197 с.
9. Еськова, О.И. Основы статистической обработки информации: пособие / О.И. Еськова, Л.П. Авдашкова, М.А. Грибовская. – Минск: Беларусь, 2011. – 175 с.: ил.
10. Гмурман, В.Е. Руководство к решению задач по теории вероятностей и математической статистике: Учеб. Пособие для студентов вузов / В.Е. Гмурман. – 7-е изд., доп. – М.: Высш. шк., 2003. – 405 с.: ил.
11. Булдык, Г.М. Руководство к решению задач и упражнений по теории вероятностей и математической статистике: Для практической и самостоятельной работы студентов экономических специальностей. – Минск: ФУАинформ, 2009. – 228 с.
12. Лунгу, К.Н. Высшая математика. Руководство к решению задач. Ч. 2 / К.Н. Лунгу, Е.В. Макаров – М.: ФИЗМАТЛИТ, 2007. – 384 с.

13. 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. – Минск, БНТУ, 2009. – 70 с.
14. https://function-x.ru/statistics_dispersion_analysis.html
15. https://studfile.net/preview/5711189/
16. https://studfile.net/preview/5520528/page:35/