МИНИСТЕРСТВО ОБРАЗОВАНИЯ И НАУКИ РФ
АЛТАЙСКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

Е.П. Петров

СТАТИСТИЧЕСКИЙ АНАЛИЗ ДАННЫХ

Учебно-методическое пособие

Барнаул
Издательство
Алтайского государственного
университета
2018

Автор-составитель:
канд. физ.-мат. наук, доцент ***Е.П. Петров***

Учебно-методическое пособие рекомендовано магистрантам, обучающимся по направлению «Реклама и связи с общественностью», для практического применения во время лабораторных занятий по указанному курсу. В данном учебно-методическом издании рассматриваются вопросы практического применения статистических методов и процедур на примерах из различных областей человеческой жизнедеятельности. В качестве базового инструментального средства для статистического анализа рекомендуется использование широко распространенного программного приложения Microsoft Excel, входящего в состав пользовательского пакета Microsoft Office.

Методические указания разработаны на основе учебного пособия Борздовой Т.В. «Основы статистического анализа и обработка данных с применением Microsoft Excel». – Минск: ГИУСТ БГУ, 2011.

Подписано в печать 03.05.2018. Формат 60x84/16
Усл.-печ. л. 2,56. Тираж 100 экз. Заказ № 193
Типография Алтайского государственного университета:
656099, Барнаул, ул. Димитрова, 66

Содержание

1 Аппроксимация экспериментальных данных **4**
1.1 Одна независимая переменная 4
1.2 Несколько независимых переменных 9
1.3 Задания для самостоятельной работы 12

2 Определение выборочных характеристик **13**
2.1 Основные понятия и определения 14
2.2 Построение распределения выборочных данных в Excel 15
2.3 Основные выборочные характеристики 18
2.4 Определение основных статистических характеристик средствами Мастера функций 19
2.5 Задания для самостоятельной работы 23

3 Принятие статистических решений **23**
3.1 Построение доверительных интервалов для среднего 24
3.2 Вычисление доверительных интервалов в Excel 25
3.3 Проверка соответствия теоретическому распределению (критерий согласия хи-квадрат) 27
3.4 Использование критерия хи-квадрат в Excel 28
3.5 Анализ двух выборок – t-критерий Стьюдента (критерий различия) . . 30
3.6 Использование t-критерия Стьюдента в Excel 31
3.7 Анализ двух выборок – критерий согласия хи-квадрат 34
3.8 Задания для самостоятельной работы 36

4 Дисперсионный анализ **38**

5 Корреляционный анализ **39**
5.1 Задания для самостоятельной работы 42

6 Литература, Интернет-ресурсы **43**

1 Аппроксимация экспериментальных данных

На практике часто приходится сталкиваться с задачей о сглаживании экспериментальной зависимости или задачей аппроксимации. Рассмотрим более подробно эту задачу и каким образом она реализуется средствами Microsoft Excel.

1.1 Одна независимая переменная

В простейшем случае задача аппроксимации экспериментальных данных выглядит следующим образом.

Пусть есть какие-то данные, полученные практическим путем (в ходе эксперимента или наблюдения), которые можно представить парами чисел (x, y). Зависимость между ними $y = f(x)$ отражает следующая таблица:

x	x_1	x_2	$\ldots$	x_n
y	y_1	y_2	$\ldots$	y_n

.

На основе этих данных требуется подобрать функцию $y = \varphi(x)$, которая наилучшим образом сглаживала бы экспериментальную зависимость между переменными и по возможности точно отражала общую тенденцию зависимости между x и y, исключая погрешности измерений и случайные отклонения. Это значит, что отклонения $f(x_i) - \varphi(x_i)$ в каком-то смысле должны быть наименьшими.

Выяснить вид функции можно анализируя расположение точек (x_i, y_i) на координатной плоскости. Например, пусть точки расположены так, как показано на рис. 1.

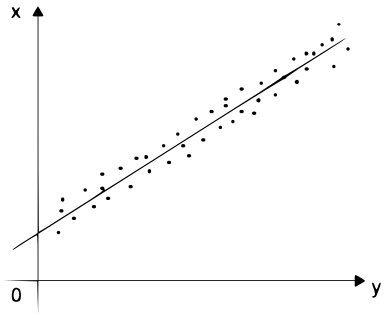


Рис. 1: прямая линия

Учитывая то, что практические данные получены с некоторой погрешностью, обусловленной неточностью измерений, необходимостью округления результатов и т. п., естественно предположить, что здесь имеет место линейная зависимость $y = ax+b$. Чтобы функция приняла конкретный вид, необходимо каким-то образом с помощью специальных формул вычислить a и b (например, с помощью так называемого метода наименьших квадратов).

Расположение экспериментальных точек в виде кривой на рис. 2 наводит на мысль, что зависимость обратно пропорциональна и функцию $y = \varphi(x)$ нужно подбирать в виде $y = a + \frac{b}{x}$. Здесь также необходимо вычислить параметры a и b. Таким образом, расположение экспериментальных точек может иметь самый различный вид, и каждому соответствует конкретный тип функции.

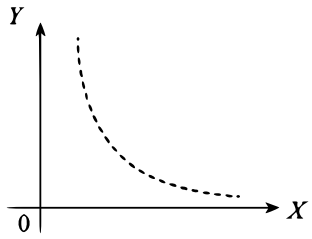


Рис. 2: гипербола

Построение эмпирической функции сводится к вычислению входящих в нее параметров так, чтобы из всех функций такого вида выбрать ту, которая лучше других описывает зависимость между изучаемыми величинами. То есть сумма квадратов разности между табличными значениями функции в некоторых точках и значениями, вычисленными по полученной формуле, должна быть минимальна.

В MS Excel аппроксимация экспериментальных данных осуществляется путем построения их графика (x — отвлеченные величины) или точечного графика (x имеет конкретные значения) с последующим подбором подходящей аппроксимирующей функции (линии тренда).

Возможны следующие варианты функций:

1. *Линейная:* $y = ax + b$. Обычно применяется в простейших случаях, когда экспериментальные данные возрастают или убывают с постоянной скоростью.

2. *Полиномиальная:* $y = a_0 + a_1x + a_2x^2 + ... + a_nx^n$ до шестого порядка включительно ($n \leq 6$), a_i — константы. Используется для описания экспериментальных данных, попеременно возрастающих и убывающих. Степень полинома определяется количеством экстремумов (максимумов или минимумов) кривой. Полином второй степени может описать только один максимум или минимум, полином третьей степени может иметь один или два экстремума, четвертой степени — не более трех экстремумов и т. д.

3. *Логарифмическая:* $y = a \ln x + b$, где a и b — константы, ln — функция натурального логарифма. Функция применяется для описания экспериментальных данных, которые вначале быстро растут или убывают, а затем постепенно стабилизируются.

4. *Степенная:* $y = bx^a$, где a и b — константы. Аппроксимация степенной функцией используется для экспериментальных данных с постоянно увеличивающейся (или убывающей) скоростью роста. Данные не должны иметь нулевых или отрицательных значений.

5. *Экспоненциальная:* $y = be^{ax}$, где a и b — константы, e — основание натурального логарифма. Применяется для описания экспериментальных данных, которые быстро растут или убывают, а затем постепенно стабилизируются. Часто ее использование вытекает из теоретических соображений.

Степень близости аппроксимации экспериментальных данных выбранной функцией оценивается коэффициентом детерминации (R^2). Таким образом, если есть несколько подходящих вариантов типов аппроксимирующих функций, можно выбрать функцию с большим коэффициентом детерминации (стремящимся к 1).

Для осуществления аппроксимации на диаграмме экспериментальных данных в случае использования пакета Microsoft Excel необходимо щелчком правой кнопки мыши вызвать контекстное меню и выбрать пункт **Добавить линию тренда**. В появившемся диалоговом окне **Линия тренда** на вкладке **Тип** выбирается вид

аппроксимирующей функции, а на вкладке **Параметры** задаются дополнительные параметры, влияющие на отображение аппроксимирующей кривой.

Пример 1. Исследовать характер изменения с течением времени уровня производства некоторой продукции и подобрать аппроксимирующую функцию, располагая следующими данными:

Год	**Производство продукции**
1997	17,1
1998	18,0
1999	18,9
2000	19,7
2001	19,7

.

Решение.

1. Для построения диаграммы прежде всего необходимо ввести данные в рабочую таблицу.

	A	B
1	**Год**	**Производство продукции**
2	1997	17,1
3	1998	18
4	1999	18,9
5	2000	19,7
6	2001	19,7

Рис. 3: рабочая таблица

2. Далее по введенным в рабочую таблицу данным необходимо построить диаграмму. Поскольку здесь необходимо показать динамику изменений производства продукции, не привязываясь к конкретному году, а от отвлеченных переменных, — выберем диаграмму **График**.

Получен график экспериментальных данных.

Рис. 4: график

3. Осуществим аппроксимацию полученной кривой полиномиальной функцией второго порядка, поскольку кривая довольно гладкая и не сильно отличается от прямой линии. Для этого указатель мыши устанавливаем на одну из точек графика и

щелкаем правой кнопкой. В появившемся контекстном меню выбираем пункт **Добавить линию тренда**. Появляется диалоговое окно **Линия тренда**.

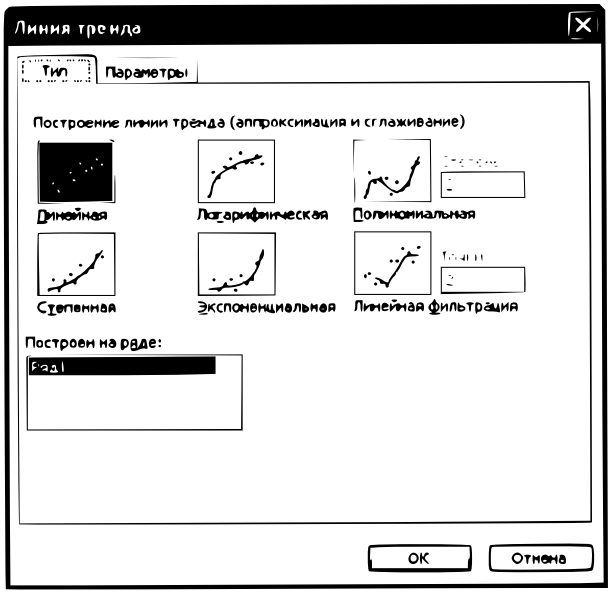


Рис. 5: линии тренда

В этом окне на вкладке **Тип** выбираем тип линии тренда -– **Полиномиальная** — и устанавливаем степень — **2**. Затем открываем вкладку **Параметры** и устанавливаем флажки в поля **показывать уравнение на диаграмме** и **поместить на диаграмму величину достоверности аппроксимации** (R^2).

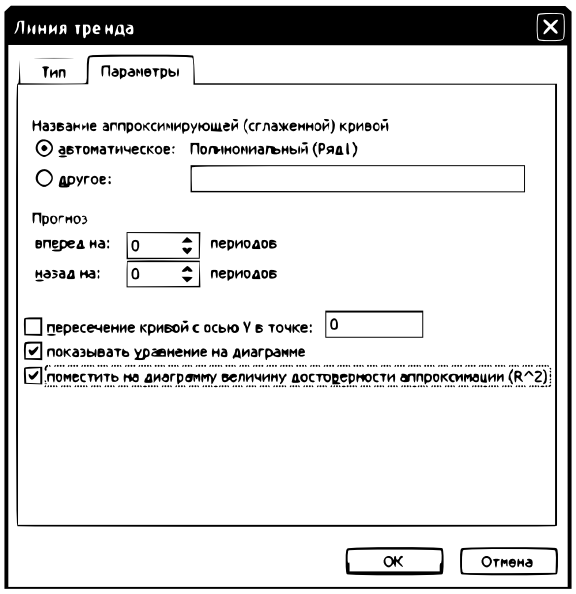


Рис. 6: параметры

После чего нужно щелкнуть по кнопке **OK**. В результате получим на диаграмме аппроксимирующую кривую.

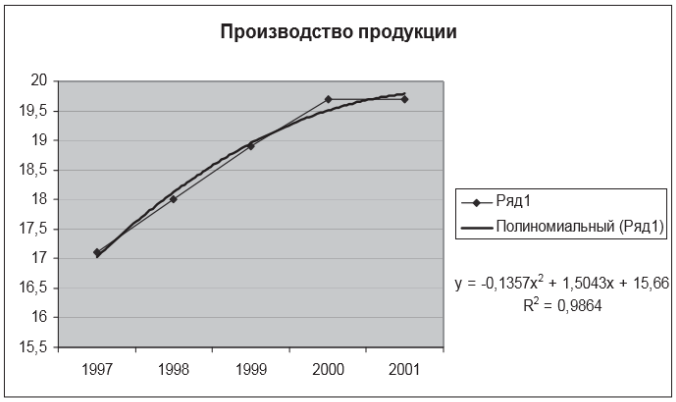


Рис. 7: аппроксимирующая кривая

Как видно из рисунка, уравнение наилучшей полиномиальной аппроксимирующей функции для некоторых отвлеченных значений выглядит как $y = -0,14x^2 + 1,5 + 15,66$. При этом точность аппроксимации достаточно высока — $R^2 = 0,986$.

4. Попробуем улучшить качество аппроксимации выбором другого типа функции (возможно, более адекватного). Здесь возможным вариантом представляется логарифмическая функция. Для этого повторяем операции п.3 за исключением того, что в окне **Линия тренда** на вкладке **Тип** выбираем тип линии тренда — **Логарифмическая**. В результате получим другой вариант аппроксимации — логарифмической кривой.

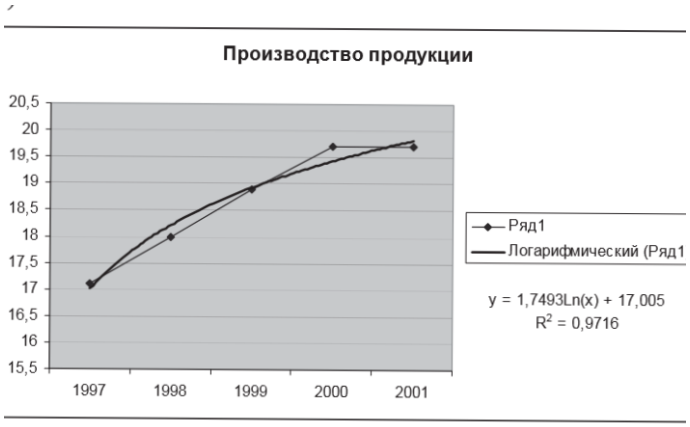


Рис. 8: логарифмическая кривая

Как можно видеть из рисунка, уравнение наилучшей логарифмической аппроксимирующей функции несколько уступает по точности аппроксимации полиномиальной кривой $R^2 = 0,9716 < 0,986$. Поэтому, если нет каких-либо теоретических соображений, то можно считать, что наилучшей аппроксимацией является аппроксимация полиномиальной функцией второй степени (из двух рассмотренных вариантов).

1.2 Несколько независимых переменных

В тех случаях, когда аппроксимируемая переменная y зависит от нескольких независимых переменных $x_1, x_2, \ldots, x_n$, т. е. $y = f(x_1, x_2, \ldots, x_n)$, подход с построением линии тренда не дает решения. Здесь могут быть использованы следующие специальные функции MS Excel:

ЛИНЕЙН и **ТЕНДЕНЦИЯ** для аппроксимации линейных функций вида: $y = a_0 + a_1x_1 + a_2x_2 + ... + a_nx_n$,

ЛГРФПРИБЛ и **РОСТ** для аппроксимации показательных функций вида: $y = a_0a_1^{x_1}a_2^{x_2}\cdots a_n^{x_n}$.

Функции **ЛИНЕЙН** и **ЛГРФПРИБЛ** служат для вычисления неизвестных коэффициентов $a_0, dots, a_n$, а также коэффициентов детерминации (R^2), значений критерия Фишера (см. подробнее далее), стандартных ошибок коэффициентов a_i и ряда других показателей.

Синтаксис:

ЛИНЕЙН(известные значения y; известные значения x; конст; статистика)

ЛГРФПРИБЛ(известные значения y; известные значения x; конст; статистика)

Здесь:

• **известные значения y** — множество наблюдаемых значений y;

• **известные значения x** — множество наблюдаемых значений $x_1, x_2, \ldots, x_n$. Причем, если массив **известные значения y** имеет один столбец, то каждый столбец массива **известные значения x** интерпретируется как отдельная переменная, а если массив **известные значения y** имеет одну строку, то тогда каждая строка массива **известные значения x** интерпретируется как отдельная переменная;

• **конст** — логическое значение, которое указывает, требуется ли, чтобы константа a_0 была равна 0 (для функции **ЛИНЕЙН**) или 1 (для функции **ЛГРФПРИБЛ**). При этом, если **конст** имеет значение **ИСТИНА** или опущено, то a_0 вычисляется обычным образом, а если **конст** имеет значение **ЛОЖЬ**, то a_0 полагается равным 0 или 1;

• **статистика** — логическое значение, которое указывает, требуется ли вычислять дополнительную статистику по регрессии: если введено значение **ИСТИНА**, то дополнительные параметры вычисляются, если **ЛОЖЬ**, то — нет.

Функции **ТЕНДЕНЦИЯ** и **РОСТ** позволяют находить точки, лежащие на аппроксимирующих кривых для значений коэффициентов $a_0, a_1, \ldots, a_n$, найденных функциями **ЛИНЕЙН** и **ЛГРФПРИБЛ**.

Синтаксис:

ТЕНДЕНЦИЯ(известные значения y; известные значения x; новые значения x; конст);

РОСТ(известные значения y; известные значения x; новые значения x; конст)

Здесь:

• **известные значения y** — множество значений y;

• **известные значения x** — множество значений x;

• **новые значения x** — те значения x, для которых необходимо определить соот-

ветствующие аппроксимирующие или предсказанные значения y. **Новые значения x** должны содержать столбец (или строку) для каждой независимой переменной, как и **известные значения x**. Если аргумент **новые значения x** опущен, то предполагается, что он совпадает с аргументом **известные значения x**;

• **конст** — логическое значение, которое указывает, требуется ли, чтобы константа a_0 была равна 0 (для функции **ТЕНДЕНЦИЯ**) или 1 (для функции **РОСТ**). При этом, если **конст** имеет значение **ИСТИНА** или опущено, то a_0 вычисляется обычным образом, а если **конст** имеет значение **ЛОЖЬ**, то a_0 полагается равным значениям 0 или 1.

Пример 2. Источник радиоактивного излучения помещен в жидкость. Датчики расположены на расстоянии (x_1) 20, 50 и 100 см от источника. Измеренная интенсивность излучения (y, мРн) проводилась через 1, 5 и 10 суток (x_2) после установки источника. Результаты измерений (y) приведены в следующей таблице:

$\mathbf{x_1 \backslash x_2}$	**1**	**5**	**10**
20	61,2	43,6	28,3
50	33,6	24,0	15,6
100	12,3	8,8	5,7

Необходимо аппроксимировать данные и найти неизвестные параметры.

Решение.

1. Введем данные в рабочую таблицу: в ячейку A1 текст x_1, в ячейку B1 — x_2, в ячейку C1 — y. В диапазон ячеек A2:A10 внесем значения x_1, в диапазон B2:B10 — значения x_2 и в диапазон C2:C10 — значения y.

	A	B	C
1	x_1	x_2	y
2	20	1	61,2
3	50	1	33,6
4	100	1	12,3
5	20	5	43,6
6	50	5	24
7	100	5	8,8
8	20	10	28,3
9	50	10	15,6
10	100	10	5,7

Рис. 9: рабочая таблица

2. Выделяем блок ячеек D1:F5 под массив результатов.

3. Поскольку уравнение для вычисления интенсивности излучения имеет степенной характер, вызываем функцию **ЛГРФПРИБЛ**.

4. Заполняем рабочие поля: **Известные значения y** — C2:C10, **Известные значения x** — A2:B10, **Статистика** — **ИСТИНА**. Нажимаем сочетание клавиш **CTRL+SHIFT+ENTER** (работа с массивом). В результате в диапазоне D1:F5 получим следующие данные:

D	E	F
0,918043	0,980162	99,70907
0,000337	3,76E-05	0,003051
0,999983	0,003722	#Н/Д
174174,7	6	#Н/Д
4,826734	8,31E-05	#Н/Д

Рис. 10: результаты анализа данных

Здесь первая строка — значения коэффициентов a_2, a_1, a_0 соответственно, вторая строка — стандартные ошибки этих коэффициентов, третья строка — коэффициент детерминации R^2 и стандартная ошибка y, четвертая строка — значение критерия Фишера и число степеней свободы и нижняя строка — сумма квадратов регрессии и остаточная сумма квадратов.

Таким образом, искомое аппроксимирующее уравнение имеет вид:
$y = 99,7 \cdot 0,98^{x_1} \cdot 0,92^{x_2}$.

Причем точность аппроксимации очень высокая — $R^2 = 0,99998$.

Пример 3. В бассейне проводится ежедневная частичная смена воды. Имеются данные семидневных наблюдений изменения уровня воды в бассейне (y) от продолжительности заполнения водой (x_1) и времени выпуска воды (x_2).

$\mathbf{x_1}$	$\mathbf{x_2}$	**y**
120	20	3,2
100	25	2,8
130	20	3,3
100	15	3,3
110	23	3,0
105	26	2,8
112	16	3,3

Необходимо найти значения уровня воды в бассейне в зависимости от длительности заполнения $x_1 \in [100; 130]$ и выпуска воды $x_2 \in [15; 25]$ с шагом $\Delta = 5$ минут.

Решение.

1. Введем данные в рабочую таблицу: в ячейку A1 — текст x_1, в ячейку B1 — x_2, в ячейку C1 — y. В диапазон ячеек A2:A8 внесем значения x_1, в диапазон B2:B8 — значения x_2 и в диапазон C2:C8 — значения y.

2. Введем значения x_1 и x_2 для получения расчетных значений в соответствии с заданием: $x_1 \in [100; 130]$ — в диапазон A10:A30, а $x_2 \in [15; 25]$ — в диапазон B10:B30.

3. Выделим блок ячеек C10:C30 под массив расчетных (предсказанных) значений y.

4. Поскольку уравнение для вычисления уровня воды линейное, вызываем функцию **ТЕНДЕНЦИЯ**.

5. Заполняем рабочие поля: **Известные значения y** — C2:C8; **Известные значения x** — A2:B8, **Новые значения x** — A10:B30. Нажимаем сочетание клавиш

Ctrl+Shift+Enter.

6. В результате в диапазоне C10:C30 получим предсказанные значения y.

	A	B	C
1	**x_1**	**x_2**	**y**
2	120	20	3,2
3	100	25	2,8
4	130	20	3,3
5	100	15	3,3
6	110	23	3
7	105	26	2,8
8	112	16	3,3
9	**x_1**	**x_2**	**y**
10	100	15	3,279046
11	105	15	3,319757
12	110	15	3,360469
13	115	15	3,40118
14	120	15	3,441891
15	125	15	3,482602
16	130	15	3,523314
17	100	20	3,044012
18	105	20	3,084723
19	110	20	3,125434
20	115	20	3,166145
21	120	20	3,206857
22	125	20	3,247568
23	130	20	3,288279
24	100	25	2,808977
25	105	25	2,849688
26	110	25	2,8904
27	115	25	2,931111
28	120	25	2,971822
29	125	25	3,012533
30	130	25	3,053245

Рис. 11: Расчетные и предсказанные значения

1.3 Задания для самостоятельной работы

1. Построить функцию, наилучшим образом отражающую данную зависимость:

x	1,0	1,5	3,0	4,5	5,0
y	1,25	1,4	1,5	1,75	2,25

.

2. В 80-е годы уровень дефицита бюджета в СССР и США складывался следующим образом:

страна	**годы**								
	1980	**1981**	**1982**	**1983**	**1984**	**1985**	**1986**	**1987**	**1988**
СССР	2,9	2,3	3,1	2,2	2,0	2,7	6,5	8,0	9,1
США	2,8	2,6	4,1	6,3	5,0	5,4	5,3	3,4	3,2

.

Построить функции, наилучшим образом отражающие зависимости дефицита бюджета от времени в обеих странах.

3. Количество вложенных в производство средств и полученная в результате прибыль соотносятся следующим образом:

x	1,6	2,0	2,5	3,0	4,0	7,0
y	8,5	9,0	11,0	13,0	22,0	70,0

.

Записать аналитическую зависимость между x и y. Проанализировать полученный ответ. Каковы перспективы предприятия? Какая будет прибыль, если вложить 10,0 единиц? Сколько надо вложить средств, чтобы получить прибыль 100,0 единиц?

4. Застройщик оценивает стоимость группы небольших офисных зданий в традиционном деловом районе. Оценку цены офисного здания в заданном районе застройщик предполагает осуществлять на основе следующих переменных: y – оценочная цена здания под офис, x_1 – общая площадь в квадратных метрах, x_2 – количество офисов, x_3 – количество входов, x_4 – время эксплуатации здания в годах. Предполагается, что существует линейная зависимость между каждой независимой переменной x_1, x_2, x_3 и x_4 и зависимой переменной y, то есть ценой здания под офис в данном районе. Застройщик наугад выбирает 11 зданий из имеющихся 1500 и получает следующие данные:

$\mathbf{x_1}$	$\mathbf{x_2}$	$\mathbf{x_3}$	$\mathbf{x_4}$	**y**
2310	2	2	20	142 000
2333	2	2	12	144 000
2356	3	1,5	33	151 000
2379	3	2	43	150 000
2402	2	3	53	139 000
2425	4	2	23	169 000
2448	2	1,5	99	126 000
2471	2	2	34	142 900
2494	3	3	23	163 000
2517	4	4	55	169 000
2540	2	3	22	149 000

.

Здесь "полвхода"(1/2) означает вход только для доставки корреспонденции. Найти параметры аппроксимирующего уравнения. С помощью функции **ТЕНДЕНЦИЯ** определить оценочную стоимость здания под офис в том же районе, которое имеет площадь 2500 квадратных метров, три офиса, два входа, зданию 25 лет.

2 Определение выборочных характеристик

В работе любого специалиста часто приходится сталкиваться с необходимостью обработки и анализа данных, полученных в результате наблюдения. Раздел математики, посвященный методам сбора, анализа и обработки статистических данных для научных и практических целей, называется математической статистикой. Математическая статистика имеет дело с массовыми явлениями. Она тесно связана с теорией вероятностей и базируется на ее математическом аппарате. Целью статистического

исследования является обнаружение и исследование соотношений между статистическими данными и их использование для изучения, прогнозирования и принятия решений. Статистические данные представляют собой данные, полученные в результате обследования большого числа объектов или явлений.

Пакет MS Excel оснащен средствами статистической обработки данных. И хотя Excel существенно уступает специализированным статистическим пакетам обработки данных, тем не менее этот раздел математики представлен в Excel наиболее полно. В него включены основные, часто используемые статистические процедуры: средства описательной статистики, критерии различия, корреляционные и другие методы. При рассмотрении применения методов обработки статистических данных ограничимся только простейшими и наиболее часто используемыми методами, реализованными в **Мастере функций** и **Пакете анализа** Excel.

2.1 Основные понятия и определения

Статистические данные представляют собой данные, полученные в результате обследования большого числа объектов или явлений.

Часть объектов исследования, определенным образом выбранная из более обширной совокупности, называется выборкой, а вся исходная совокупность, из которой взята выборка, – генеральной (основной) совокупностью.

Исследования, в которых участвуют все без исключения объекты, составляющие генеральную совокупность, называются сплошными исследованиями. Может использоваться выборочный метод, суть которого в том, что для обследования привлекается часть генеральной совокупности (выборка), но по результатам этого обследования судят о свойствах всей генеральной совокупности.

Предметом изучения в статистике являются, в частности, количественные признаки, которые представляют собой результаты подсчета или измерения.

Пусть X – некоторый признак изучаемого объекта или явления (срок службы электролампы, вес студента, диаметр шарика для подшипника и т.п.). Генеральной совокупностью является множество всех возможных значений этого признака, а результаты n наблюдений над признаком X дадут нам выборку объема n – первоначальные статистические данные, значения $x_1, x_2, \ldots, x_n$. При этом значение x_1 получено при первом наблюдении случайной величины X, x_2 – при втором наблюдении той же случайной величины и т.д.

Выборку преобразуют в вариационный ряд, располагая результаты наблюдений в порядке возрастания. Каждый член x_i вариационного ряда называется вариантой. Если варианта x_i появилась m раз, то число m называют частотой, а ее отношение к объему выборки $\frac{m}{n}$ – относительной частотой.

Соответствие между вариантами и их частотами принято называть эмпирическим распределением (распределением выборочных данных).

Конечной целью изучения выборочной совокупности всегда является получение информации о генеральной совокупности. Поэтому естественно стремиться сделать выборку так, чтобы она наилучшим образом представляла всю генеральную совокупность, то есть была бы репрезентативной или представительной. Для получения репрезентативной выборки необходимо четко определять, что понимается под генеральной совокупностью. Ее состав и численность зависят от объектов и целей проводимого исследования. Например, если мы хотим получить данные о поступающих во

все вузы города, то абитуриенты данного института есть выборка из более широкой генеральной совокупности — всех абитуриентов вузов города — и эта выборка не обязательно будет являться представительной. В тех случаях, когда генеральная совокупность недостаточно известна, обычно не удается предложить лучшего способа получения представительной выборки, чем случайный выбор. При этом случайная выборка формируется случайным отбором: из генеральной совокупности наудачу извлекается по одному объекту.

2.2 Построение распределения выборочных данных в Excel

В Excel для построения распределения выборочных данных используются специальная функция **ЧАСТОТА** и процедура **Пакета анализа Гистограмма**. Функция **ЧАСТОТА** вычисляет частоты появления случайной величины в интервалах значений и выводит их как массив чисел. Функция задается в качестве формулы массива.

Синтаксис:

ЧАСТОТА (массив данных; массив карманов)

Здесь:

• **массив данных** — это массив или ссылка на множество данных, для которых вычисляются частоты;

• **массив карманов** — это массив или ссылка на множество интервалов, в которые группируются значения аргумента массив данных.

Отметим, что количество элементов в возвращаемом массиве на единицу больше числа элементов в **массив карманов**. Дополнительный элемент в возвращаемом массиве содержит количество значений, больших, чем максимальное значение в интервалах.

Процедура **Гистограмма** используется для вычисления выборочных и интегральных частот попадания данных в указанные интервалы значений. Процедура выводит результаты в виде таблицы и гистограммы. Параметры диалогового окна **Гистограмма** представлены на рис. :

• во **Входной диапазон** вводится диапазон исследуемых данных;

• в поле **Интервал карманов** (необязательный параметр) может вводиться диапазон ячеек или необязательный набор граничных значений, определяющих выбранные интервалы (карманы). Эти значения должны быть введены в возрастающем порядке. В MS Excel вычисляется число попаданий данных между началом интервала и соседним большим по порядку. При этом включаются значения на нижней границе интервала и не включаются значения на верхней границе. Если диапазон карманов не был введен, то набор интервалов, равномерно распределенных между минимальным и максимальным значениями данных, будет создан автоматически;

• рабочее поле **Выходной диапазон** предназначено для ввода ссылки на левую верхнюю ячейку выходного диапазона. Размер выходного диапазона будет определен автоматически;

• переключатель **Интегральный процент** позволяет установить режим генерации интегральных процентных отношений и включения в гистограмму графика интегральных процентов;

• переключатель **Вывод графика** позволяет установить режим автоматического создания встроенной диаграммы на листе, содержащем выходной диапазон.

Пример 4. Построить эмпирическое распределение веса студентов в килограммах для следующей выборки: 64, 57, 63, 62, 58, 61, 63, 60, 60, 61, 65, 62, 62, 60, 64, 61, 59, 59, 63, 61, 62, 58, 58, 63, 61, 59, 62, 60, 60, 58, 61, 60, 63, 63, 58, 60, 59, 60, 59, 61, 62, 62, 63, 57, 61, 58, 60, 64, 60, 59, 61, 64, 62, 59, 65.

Решение.

1. В ячейку A1 введите слово *Наблюдения*, а в диапазон A2:E12 – значения веса студентов.

2. Выберите ширину интервала 1 кг. Тогда при крайних значениях веса 57 кг и 65 кг получится 9 интервалов. В ячейки G1 и G2 введите названия интервалов Вес и кг, соответственно. В диапазон G4:G12 введите граничные значения интервалов (57, 58, 59, 60, 61, 62, 63, 64, 65).

3. Введите заголовки создаваемой таблицы: в ячейки H1:H2 – *Абсолютные частоты*, в ячейки I1:I2 – *Относительные частоты*, в ячейки J1:J2 – *Накопленные частоты*.

4. Заполните столбец абсолютных частот. Для этого выделите для них блок ячеек H4:H12 (используемая функция **ЧАСТОТА** задается в виде формулы массива). Выполните функцию **ЧАСТОТА**. Для этого выберите ее из категории **Статистические Мастера функций**. В поле **Массив данных** введите диапазон данных наблюдений (A2:E12). В рабочее поле **Двоичный массив** введите диапазон интервалов (G4:G12). Последовательно нажмите комбинацию клавиш Ctrl+Shift+Enter. В столбце H4:H12 появится массив абсолютных частот.

5. В ячейке H13 найдите общее количество наблюдений (оно равно числу 55).

6. Заполните столбец относительных частот. В ячейку I4 введите формулу для вычисления относительной частоты: = H4/13. Нажмите клавишу Enter. Протягиванием скопируйте введенную формулу в диапазон I5:I12. Получим массив относительных частот.

7. Заполните столбец накопленных частот. В ячейку J4 скопируйте значение относительной частоты из ячейки I4 (0,036364). В ячейку J5 введите формулу: = J4 + + I5. Нажмите клавишу Enter. Протягиванием скопируйте введенную формулу в диапазон J6:J12. Получим массив накопленных частот.

8. В результате после форматирования получим следующую таблицу:

	A	B	C	D	E	F	G	H	I	J
1	Наблюдения						вес	абсолютные	относительные	накопленные
2	64	62	58	63	61		кг	частоты	частоты	частоты
3	57	62	63	58	58					
4	63	60	61	60	60		57	2	0,036363636	0,036363636
5	62	64	59	59	64		58	6	0,109090909	0,145454545
6	58	61	62	60	60		59	7	0,127272727	0,272727273
7	61	59	60	59	59		60	10	0,181818182	0,454545455
8	63	59	60	61	61		61	9	0,163636364	0,618181818
9	60	63	58	62	64		62	8	0,145454545	0,763636364
10	60	61	61	62	62		63	7	0,127272727	0,890909091
11	61	62	60	63	59		64	4	0,072727273	0,963636364
12	65	58	63	57	65		65	2	0,036363636	1
13								55		

Рис. 12: Частоты

9. Постройте диаграмму относительных и накопленных частот. Щелчком указателя мыши по кнопке на панели инструментов вызовите **Мастер диаграмм**. В появившемся диалоговом окне выберите вкладку **Нестандартные** и тип диаграммы **График/гистограмма2**. После нажатия кнопки **Далее** укажите диапазон дан-

ных – I4:J12. Проверьте положение переключателя **Ряды в: столбцах**. Выберите вкладку **Ряд** и с помощью мыши введите в рабочее поле **Подписи оси X** диапазон подписей оси X: G4.G12. Нажав кнопку **Далее**, введите названия осей X и Y : в рабочее поле **Ось X (категорий)** – *Вес*; **Ось Y (значений)** – *Относит. частота*; **Вторая ось Y (значений)** – *Накоплен. частота*. Нажмите кнопку **Готово**. После минимального редактирования диаграмма будет вид:

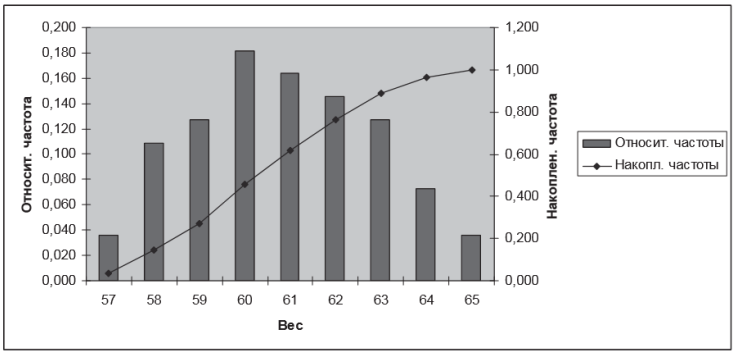


Рис. 13: Диаграмма частот

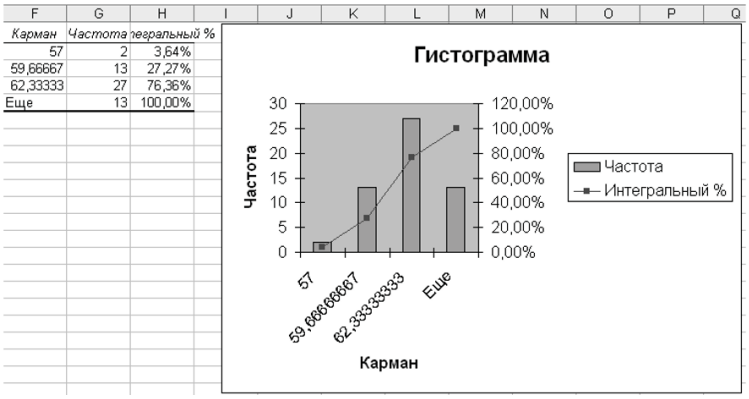
Пример 5. Для данных из примера 4 построить эмпирические распределения, воспользовавшись процедурой **Гистограмма**.

Решение.

1. В ячейку A1 введите слово *Наблюдения*, а в диапазон A2:E12 – значения веса студентов.

2. Для вызова процедуры **Гистограмма** выберите из меню **Сервис** подпункт **Анализ данных** и в открывшемся окне в поле **Инструменты анализа** укажите процедуру **Гистограмма**.

3. В появившемся окне **Гистограмма** заполните рабочие поля: во **Входной диапазон** введите диапазон исследуемых данных (A2:E12); в **Выходной диапазон** – ссылку на левую верхнюю ячейку выходного диапазона (F1). Установите переключатели в положение **Интегральный процент** и **Вывод графика**. После этого нажмите кнопку **OK**. В результате появляется таблица и диаграмма:



Карман	Частота	Интегральный %
57	2	3,64%
59,66667	13	27,27%
62,33333	27	76,36%
Еще	13	100,00%

Рис. 14: Таблица и диаграмма

2.3 Основные выборочные характеристики

Среди выборочных характеристик выделяют показатели, относящиеся к центру распределения (меры положения), показатели рассеяния вариант (меры рассеяния) и меры формы распределения. К показателям, характеризующим центр распределения, относят различные виды средних (арифметическое, геометрическое и т. п.), а также моду и медиану.

Простейшим показателем, характеризующим центр выборки, является мода.

Мода — это элемент выборки с наиболее часто встречающимся значением (наиболее вероятная величина). Средним значением выборки, или выборочным аналогом математического ожидания, называется величина

$$\overline{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

Иначе говоря, среднее значение — это центр выборки, вокруг которого группируются элементы выборки. При увеличении числа наблюдений среднее приближается к математическому ожиданию. Среднее значение обозначается также буквой M. Выборочная медиана – это число, которое является серединой выборки, то есть половина чисел имеет значения большие, чем медиана, а половина чисел имеет значения меньшие, чем медиана. Для нахождения медианы обычно выборку ранжируют – располагают элементы в порядке возрастания. Если количество членов ранжированного ряда нечетное, медианой является значение ряда, которое расположено посередине, то есть элемент с номером $\frac{n+1}{2}$. Если число членов ряда четное, то медиана равна среднему значению членов ряда с номерами $\frac{n}{2}$ и $(\frac{n}{2}+1)$.

Основными показателями рассеяния вариант являются интервал, дисперсия выборки, стандартное отклонение и стандартная ошибка.

Интервал (амплитуда, вариационный размах) –это разница между максимальным и минимальным значениями элементов выборки. Интервал является простейшей и наименее надежной мерой вариации или рассеяния элементов в выборке.

Более точно отражают рассеяние показатели, учитывающие не только крайние, но и все значения элементов выборки.

Дисперсией выборки, или выборочным аналогом дисперсии, называется величина

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \overline{x})^2.$$

Дисперсия выборки — это параметр, характеризующий степень разброса элементов выборки относительно среднего значения. Чем больше дисперсия, тем дальше отклоняются значения элементов выборки от среднего значения.

Выборочным стандартным отклонением (среднее квадратичное отклонение) называется величина

$$s = \sqrt{s^2}.$$

Этот параметр также характеризует степень разброса элементов выборки относительно среднего значения. Чем больше среднее квадратичное отклонение, тем дальше отклоняются значения элементов выборки от среднего значения. Параметр аналогичен дисперсии и используется в тех случаях, когда необходимо, чтобы показатель разброса случайной величины выражался в тех же единицах, что и среднее

значение этой случайной величины. Часто выборочное стандартное отклонение обозначают буквой σ (сигма).

Стандартная ошибка или ошибка среднего находится из выражения

$$m = \frac{s}{\sqrt{n}}.$$

Стандартная ошибка — это параметр, характеризующий степень возможного отклонения среднего значения, полученного на исследуемой ограниченной выборке, от истинного среднего значения, полученного на всей совокупности элементов. С помощью стандартной ошибки задается так называемый доверительный интервал. 95-процентный доверительный интервал, равный $x \pm 2m$, обозначает диапазон, в который с вероятностью $p = 0,95$ (при достаточно большом числе наблюдений $n > 30$) попадает среднее генеральной совокупности $M(X)$.

Показателями, характеризующими форму распределения, являются выборочные характеристики эксцесс и асимметрия.

Эксцесс – это степень выраженности «хвостов» распределения, то есть частоты появления удаленных от среднего значений.

Асимметрия – величина, характеризующая несимметричность распределения элементов выборки относительно среднего значения. Принимает значения от -1 до 1. В случае симметричного распределения асимметрия равна 0.

Часто значения асимметрии и эксцесса используют для проверки гипотезы о том, что данные (выборка) принадлежат к определенному теоретическому распределению, в частности, нормальному распределению. Для нормального распределения асимметрия равна нулю, а эксцесс – трем.

2.4 Определение основных статистических характеристик средствами Мастера функций

В **Мастере функций** Excel имеется ряд специальных функций, предназначенных для вычисления выборочных характеристик. Прежде всего, это функции, характеризующие центр распределения.

Функция **СРЗНАЧ** вычисляет среднее арифметическое из нескольких массивов (аргументов) чисел.

Функция **СРГАРМ** позволяет получить среднее гармоническое множества данных. Среднее гармоническое – это величина, обратная к среднему арифметическому обратных величин. Например: **СРГАРМ**(10;14;5;6;10;12;13) равняется $8,317$.

Функция **СРГЕОМ** вычисляет среднее геометрическое значений массива положительных чисел. Функцию **СРГЕОМ** можно использовать для вычисления средних показателей динамического ряда. Например: **СРГЕОМ**(10;14;5;6;10;12;13) равняется $9,414$.

Функция **МЕДИАНА** позволяет получать медиану заданной выборки. Медиана – это элемент выборки, число элементов выборки со значениями больше которого и меньше которого равно. Например: **МЕДИАНА**(10;14;5;6;10;12;13) равняется 10.

Функция **МОДА** вычисляет наиболее часто встречающееся значение в выборке. Например: **МОДА**(10;14;5;6;10;12;13) равняется 10.

К специальным функциям, вычисляющим выборочные характеристики, характеризующие рассеяние вариант, относятся ДИСП, СТАНДОТКЛОН.

Функция **ДИСП** позволяет оценить дисперсию по выборочным данным. Например: **ДИСП**(10;14;5;6;10;12;13) равняется $11,667$.

Функция **СТАНДОТКЛОН** вычисляет стандартное отклонение. Например: **СТАНДОТКЛОН**(10;14;5;6;10;12;13) равняется $3,416$.

Форму эмпирического распределения позволяют оценить специальные функции **ЭКСЦЕСС** и **СКОС**.

Функция **ЭКСЦЕСС** вычисляет оценку эксцесса по выборочным данным. Например: **ЭКСЦЕСС**10;14;5;6;10;12;13) равняется $-1,169$.

Функция **СКОС** позволяет оценить асимметрию выборочного распределения. Например: **СКОС**(10;14;5;6;10;12;13) равняется $-0,527$.

Пример 6. Рассматриваются ежемесячные количества реализованных турфирмой путевок за периоды до и после начала активной рекламной компании. Ниже приведены количества реализованных путевок по месяцам. Требуется найти средние значения и стандартные отклонения этих данных.

С рекламой	162	156	144	137	125	145	151
Без рекламы	135	126	115	140	121	112	130

.

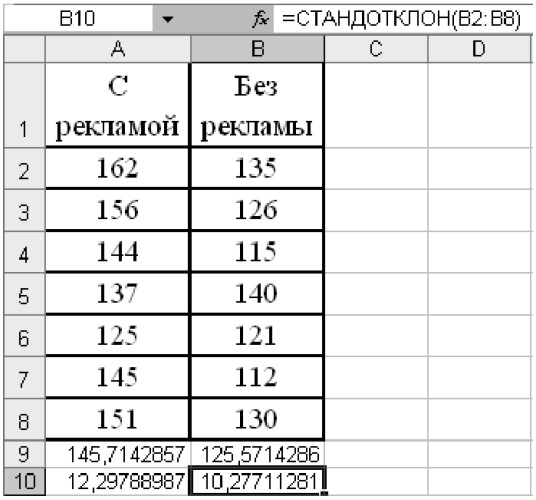
Решение.

1. Для проведения статистического анализа прежде всего необходимо ввести данные в рабочую таблицу, как показано ниже.

	A	B
1	С рекламой	Без рекламы
2	162	135
3	156	126
4	144	115
5	137	140
6	125	121
7	145	112
8	151	130

2. При статистическом анализе необходимо определить характеристики выборки, при этом важнейшей характеристикой является среднее значение. Для определения среднего значения в контрольной группе необходимо установить табличный курсор в свободную ячейку (например, A9 и B9) и вызвать функцию **СРЗНАЧ** для диапазона значений A2:A8 и B2:B8. В соответствующих ячейках получим значения $145,714$ и $125,571$.

3. Следующей по важности характеристикой выборки является мера разброса элементов выборки от среднего значения. Такой мерой является среднее квадратичное или стандартное отклонение. Для определения стандартного отклонения в контрольной группе необходимо установить табличный курсор в свободную ячейку (например, A10 и B10) и вызвать функцию **СТАНДОТКЛОН**. В соответствующих ячейках получим значения $12,298$ и $10,277$. Существует правило, согласно которому данные должны лежать в диапазоне $M \pm 3\sigma$ (в примере $145,7 \pm 36,9$).



B10 =СТАНДОТКЛОН(B2:B8)

	A	B	C	D
1	С рекламой	Без рекламы		
2	162	135		
3	156	126		
4	144	115		
5	137	140		
6	125	121		
7	145	112		
8	151	130		
9	145,7142857	125,5714286		
10	12,29788987	10,27711281		

В пакете Excel помимо **Мастера функций** имеется набор более мощных инструментов для работы с несколькими выборками и углубленного анализа данных, называемый **Пакет анализа**, который может быть использован для решения задач статистической обработки выборочных данных. Для установки **Пакета анализа** в Excel выполните следующее: в меню **Сервис** выберите команду **Надстройки**; в появившемся списке установите флажок **Пакет анализа**.

Последовательность обработки данных.

Для использования статистического пакета анализа данных необходимо:

- выполнить команду **Сервис - Анализ данных**;
- выбрать необходимую строку в появившемся списке **Инструменты анализа**;
- ввести входной и выходной диапазоны и выбрать необходимые параметры.

Пример 7. Рассматривается зарплата основных групп работников гостиницы: администрации, обслуживающего персонала и работников ресторана. Были получены следующие данные:

Администрация	**Персонал**	**Ресторан**
4500	2100	3200
4000	2100	3000
3700	2000	2500
3000	2000	2000
2500	2000	1900
	1900	1800
	1800	
	1800	

.

Необходимо определить основные статистические характеристики в группах данных.

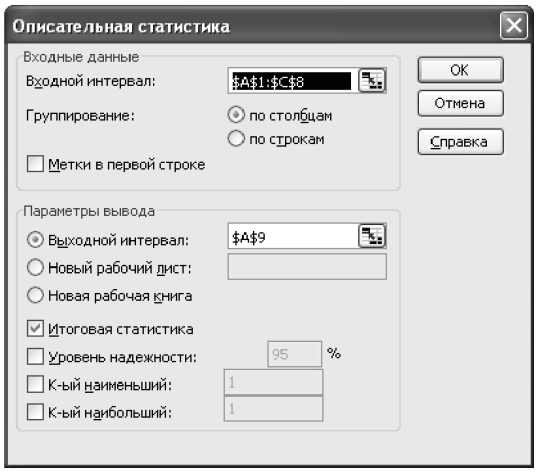
Решение.

1. Для использования инструментов анализа исследуемые данные следует представить в виде таблицы, где столбцами являются соответствующие показатели. Значения зарплат сотрудников администрации введите в диапазон A1:A5, обслуживающего персонала – в диапазон B1:B8 и т. д. В результате получится следующая таблица:

	A	B	C
1	4500	2100	3200
2	4000	2100	3000
3	3700	2000	2500
4	3000	2000	2000
5	2500	2000	1900
6		1900	1800
7		1800	
8		1800	

2. Далее необходимо провести элементарную статистическую обработку. Для этого выполните команду **Сервис – Анализ данных**. Затем в появившемся списке **Инструменты анализа** выберите строку **Описательная статистика**.

3. В появившемся диалоговом окне (рис. 3.7) в рабочем поле **Входной интервал** укажите входной диапазон – A1:C8. Активировав переключателем рабочее поле **Выходной интервал**, укажите выходной диапазон – ячейку A9. В разделе **Группировка** переключатель установите в положение **по столбцам**. Установите флажок в поле **Итоговая статистика** и нажмите кнопку **OK**.



В результате анализа в указанном выходном диапазоне для каждого столбца данных получим соответствующие результаты.

9	*Столбец1*		*Столбец2*		*Столбец3*	
10						
11	Среднее	3540	Среднее	1962,5	Среднее	2400
12	Стандартная ошибка	356,8089375	Стандартная ошибка	41,99277	Стандартная ошибка	243,6843
13	Медиана	3700	Медиана	2000	Медиана	2250
14	Мода	#Н/Д	Мода	2000	Мода	#Н/Д
15	Стандартное отклонение	795,6129712	Стандартное отклонение	118,7735	Стандартное отклонение	596,6574
16	Дисперсия выборки	633000	Дисперсия выборки	14107,14	Дисперсия выборки	356000
17	Эксцесс	-1,29384635	Эксцесс	-1,22929	Эксцесс	-2,06887
18	Асимметричность	-0,245024547	Асимметричность	-0,39433	Асимметричность	0,457606
19	Интервал	2000	Интервал	300	Интервал	1400
20	Минимум	2500	Минимум	1800	Минимум	1800
21	Максимум	4500	Максимум	2100	Максимум	3200
22	Сумма	17700	Сумма	15700	Сумма	14400
23	Счет	5	Счет	8	Счет	6

2.5 Задания для самостоятельной работы

1. Построить эмпирические функции распределения (относительные и накопленные частоты) для роста (в см) группы из 20 мужчин: $181, 169, 178, 178, 171, 179, 172, 181, 179, 168, 174, 167, 169, 171, 179, 181, 181, 183, 172, 176$.

2. Найти распределение по абсолютным частотам для следующих результатов тестирования в баллах: $79, 85, 78, 85, 83, 81, 95, 8897$.

3. Построить эмпирические функции распределения (абсолютные и накопленные частоты) успеваемости в группе из 20 студентов:
$4, 4, 5, 3, 4, 5, 4, 5, 3, 5, 3, 3, 5, 4, 5, 4, 3, 5, 3, 5$.

4. Найти среднее значение и стандартное отклонение результатов бега на дистанцию 100 м у группы студентов: $12,8; 13,2; 13,0; 12,9; 13,5; 13,1$.

5. Найти выборочные среднее, медиану, моду, дисперсию и стандартное отклонение для следующей выборки: $26, 35, 29, 27, 33, 35, 30, 33, 31, 29$.

6. Найти наиболее популярный туристический маршрут из четырех реализуемых фирмой (моду), если за неделю последовательно были реализованы следующие маршруты (приводятся номера маршрутов):
$1, 3, 3, 2, 1, 1, 4, 4, 2, 4, 1, 3, 2, 4, 1, 4, 4, 3, 1, 2, 3, 4, 1, 1, 3$.

7. В рабочей зоне производились замеры концентрации вредного вещества. Получен ряд значений (в мг/м3): $12, 16, 15, 14, 10, 20, 16, 14, 18, 14, 15, 17, 23, 16$. Необходимо определить основные выборочные характеристики.

3 Принятие статистических решений

сс Статистическая гипотеза – это предположение о виде или отдельных параметрах распределения вероятностей, которое подлежит проверке на имеющихся данных.

Проверка статистических гипотез – это процесс формирования решения о возможности принять или отвергнуть утверждение (гипотезу), основанный на информации, полученной из анализа выборки. Методы проверки гипотез называются критериями.

В большинстве случаев рассматривают так называемую нулевую гипотезу (нуль-гипотезу H_0), состоящую в том, что все события произошли случайно, естественным образом. Альтернативная гипотеза (H_1) состоит в том, что события случайным образом произойти не могли, и имело место воздействие некого фактора. Обычно нулевая гипотеза формулируется таким образом, чтобы на основании эксперимента или наблюдений ее можно было отвергнуть с заранее заданной вероятностью ошибки α. Эта заранее заданная вероятность ошибки называется уровнем значимости.

Уровень значимости – максимальное значение вероятности появления события, при котором событие считается практически невозможным. В статистике наибольшее распространение получил уровень значимости, равный $\alpha = 0,05$. Поэтому, если вероятность, с которой интересующее событие может произойти случайным образом $p < 0,05$, то принято считать это событие маловероятным, и если оно все же произошло, то это не было случайным. В наиболее ответственных случаях, когда

требуется особая уверенность в достоверности полученных результатов, надежности выводов, уровень значимости принимают равным $\alpha = 0,01$ или даже $\alpha = 0,001$.

Величину P, равную $(1-\alpha)$, называют доверительной вероятностью (уровнем надежности), то есть вероятностью, признанной достаточной для того, чтобы уверенно судить о принятом статистическом решении. Соответственно, в качестве доверительных вероятностей выбирают значения $0,95$, $0,99$ или $0,999$.

Интервал, в котором с заданной доверительной вероятностью $P = 1-\alpha$ находится оцениваемый параметр, называется доверительным интервалом. В соответствии с доверительными вероятностями на практике используются 95-, 99-, 99,9-процентные доверительные интервалы. Граничные точки доверительного интервала называют доверительными пределами.

Выбор того или иного уровня значимости, выше которого результаты отвергаются как статистически не подтвержденные, в общем случае является произвольным. Окончательное решение зависит от исследователя, традиций и накопленного практического опыта в данной области исследований.

3.1 Построение доверительных интервалов для среднего

Еще одной важной задачей, возникающей при анализе одной выборки, является сравнение выборочного среднего арифметического со средним значением генеральной совокупности. Эта задача решается с помощью статистических критериев. При этом выясняется, значимо ли отличие выборочного среднего значения от среднего значения генеральной совокупности, из которой предположительно взята выборка, или наблюдаемое различие является случайным.

Действительно, средние значения, получаемые по выборочным данным, обычно не совпадают с генеральным средним (математическим ожиданием). В связи с этим возникает вопрос: можно ли по результатам выборочной оценки судить о свойствах всей генеральной совокупности?

Поскольку каждую оценку, полученную в отдельной выборке, можно рассматривать как случайную величину, то при увеличении числа выборок распределение отдельных оценок будет принимать характер нормального распределения. Это значит, что в случае средних арифметических значения выборочных средних относительно генерального среднего распределяются по нормальному закону. То есть так же, как относительные отклонения нормально распределенных вариант от среднего арифметического выборки.

Отсюда, в частности, следует, что $68,3\%$ всех выборочных средних находятся в пределах $\Delta = M \pm m$, где Δ – предельная ошибка выборки, M – среднее выборочное, m – стандартное отклонение среднего значения. Иными словами, имеется вероятность $0,683$, что выборочное среднее отличается от генерального не более, чем на $\pm m$. Здесь $0,683$ – доверительная вероятность, $(1 - 0,683) = 0,317$ – уровень значимости α, $\Delta = M \pm m$ – 68%-доверительный интервал.

Для принятой в большинстве исследований доверительной вероятности $0,95$ доверительный интервал для средних при достаточно большом числе наблюдений $(n > 30)$ примерно равен $\pm 2m$. При доверительной вероятности $0,99$ доверительный интервал составит примерно $\pm 3m$.

3.2 Вычисление доверительных интервалов в Excel

В MS Excel для более точного вычисления границ доверительного интервала и при числе элементов в выборке $n < 30$ можно воспользоваться функцией **ДОВЕРИТ** или процедурой **Описательная статистика**.

Функция **ДОВЕРИТ(альфа; станд откл; размер)** определяет полуширину доверительного интервала и содержит следующие параметры:

- **альфа** — уровень значимости, используемый для вычисления доверительной вероятности. Доверительная вероятность равняется $100 \cdot (1 \breve{} \alpha)$ процентам, или, другими словами, $\alpha = 0,05$ означает 95-процентный уровень доверительной вероятности;

- **станд откл** — стандартное отклонение генеральной совокупности для интервала данных, предполагается известным;

- **размер** — это размер выборки.

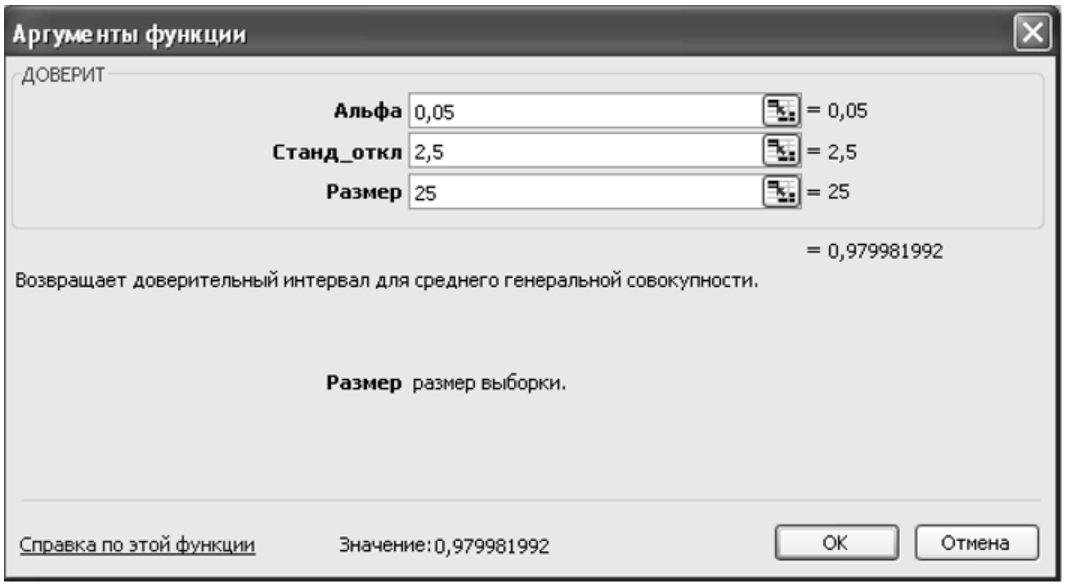
Пример 8. Найти границы 95-процентного доверительного интервала для среднего значения, если у 25 телефонных аккумуляторов среднее время разряда в режиме ожидания составило 140 часов, а стандартное отклонение – $2,5$ часа.

Решение.

1. Откройте новую рабочую таблицу. Установите табличный курсор в ячейку A1.

2. Для определения границ доверительного интервала необходимо на панели инструментов **Стандартная** нажать кнопку **Вставка функции** (f_x). В появившемся диалоговом окне **Мастера функций** выберите категорию **Статистические** и функцию **ДОВЕРИТ**, после чего нажмите кнопку **ОК**.

3. В рабочие поля появившегося диалогового окна функции **ДОВЕРИТ** с клавиатуры введите условия задачи: **Альфа** — $0,05$; **Станд откл** — $2,5$; **Размер** — 25. Нажмите кнопку **ОК**.



4. В ячейке A1 появится полуширина 95-процентного доверительного интервала для среднего значения выборки — $0,979981$. Другими словами, с 95-процентным уровнем надежности можно утверждать, что средняя продолжительность разряда аккумулятора составляет $140 \pm 0,979981$ часа или от $139,02$ до $140,98$ часа.

Пример 9. Пусть имеется выборка, содержащая числовые значения: $13, 15, 17, 19, 22, 25, 19$. Необходимо определить границы 95-процентного доверительного интервала для среднего значения и для нахождения "выскакивающей"варианты (для нахождения доверительных границ для "выскакивающей"варианты необходимо полученный доверительный интервал умножить на $\sqrt{n}$).

Решение.

1. В диапазон A1:A7 введите исходный ряд чисел.

2. Далее вызовите процедуру **Описательная статистика**. Для этого выполните команду **Сервис — Анализ данных**. Затем в появившемся списке **Инструменты анализа** выберите строку **Описательная статистика**.

3. В появившемся диалоговом окне в рабочем поле **Входной интервал** укажите входной диапазон — A1:A7. Переключателем активизируйте **Выходной интервал** и укажите выходной диапазон — ячейку B1. В разделе **Группировка** переключатель установите в положение **по столбцам**. Установите флажок **Уровень надежности** и справа от него задайте (%) — 95. Затем нажмите кнопку **OK**.

4. В результате анализа в указанном выходном диапазоне для доверительной вероятности $0,95$ получаем значения доверительного интервала.

	A	B	C
1	13	*Столбец1*	
2	15		
3	17	Уровень надежности(95,0%)	3,770269
4	19		
5	22		
6	25		
7	19		

Уровень надежности — это половина доверительного интервала для генерального среднего арифметического. Из полученного результата следует, что с вероятностью $0,95$ среднее арифметическое для генеральной совокупности находится в интервале $18,571 \pm 3,77$. Здесь $18,571$ — выборочное среднее M для рассматриваемого примера, которое находится обычно процедурой **Описательная статистика** одновременно с доверительным интервалом.

5. Для нахождения доверительных границ для «выскакивающей» варианты необходимо полученный выше доверительный интервал умножить на $\sqrt{n}$ (в примере — $\sqrt{7}$, то есть $3,77 \cdot \sqrt{7} = 9,975$). В Excel это можно выполнить следующим образом: ввести, например, в ячейку C4 формулу =C3*Корень(7). В результате получим в ячейке C4 значение доверительного интервала — $9,975$.

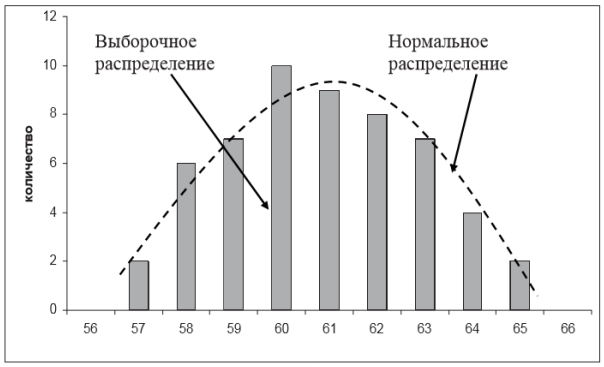
Таким образом, варианта, попадающая в интервал $18,571 \pm 9,975$, считается принадлежащей данной совокупности с вероятностью $0,95$. Выходящая за эти границы может быть отброшена с уровнем значимости $\alpha = 0,05$.

3.3 Проверка соответствия теоретическому распределению (критерий согласия хи-квадрат)

Следующей задачей, возникающей при анализе одной выборки, является оценка меры соответствия (расхождения) полученных эмпирических данных и каких-либо теоретических распределений. Это связано с тем, что в большинстве случаев при решении реальных задач закон распределения и его параметры неизвестны. В то же время применяемые статистические методы в качестве предпосылок часто требуют определенного закона распределения.

Наиболее часто проверяется предположение о нормальном распределении генеральной совокупности, поскольку большинство статистических процедур ориентировано на выборки, полученные из нормально распределенной генеральной совокупности.

Для оценки соответствия имеющихся экспериментальных данных нормальному закону распределения обычно используют графический метод, выборочные параметры формы распределения и критерии согласия. Графический метод позволяет давать ориентировочную оценку расхождения или совпадений распределений.



При большом числе наблюдений ($n > 100$) неплохие результаты дает вычисление выборочных параметров формы распределения: эксцесса и асимметрии. Принято говорить, что предположение о нормальности распределения не противоречит имеющимся данным, если асимметрия близка к нулю, то есть лежит в диапазоне от $-0,2$ до $0,2$, а эксцесс — от 2 до 4.

Наиболее убедительные результаты дает использование критериев согласия. Критериями согласия называют статистические критерии, предназначенные для проверки согласия опытных данных и теоретической модели. Здесь нулевая гипотеза H0 представляет собой утверждение о том, что распределение генеральной совокупности, из которой получена выборка, не отличается от нормального. Среди критериев согласия большое распространение получил непараметрический критерий χ^2 (хи-квадрат). Он основан на сравнении эмпирических частот интервалов группировки с теоретическими (ожидаемыми) частотами, рассчитанными по формулам нормального распределения.

Отметим, что сколько-нибудь уверенно о нормальности закона распределения можно судить, если имеется не менее 50 результатов наблюдений. В случаях меньшего числа данных можно говорить только о том, что данные не противоречат нормальному закону, и в этом случае обычно используют графические методы оценки соответствия. При большем числе наблюдений целесообразно совместное использо-

вание графических и статистических (например, тест хи-квадрат или аналогичные) методов оценки, естественно дополняющих друг друга.

3.4 Использование критерия хи-квадрат в Excel

Для применения критерия желательно, чтобы объем выборки $n > 40$, выборочные данные были сгруппированы в интервальный ряд с числом интервалов не менее 7, а в каждом интервале находилось не менее 5 наблюдений (частот).

Отметим, что сравниваться должны именно абсолютные частоты, а не относительные. При этом, как и любой другой статистический критерий, критерий хи-квадрат не доказывает справедливость нулевой гипотезы (соответствие эмпирического распределения нормальному), а лишь может позволить ее отвергнуть с определенной вероятностью (уровнем значимости).

В MS Excel критерий хи-квадрат реализован в функции **ХИ2ТЕСТ**. Функция **ХИ2ТЕСТ** вычисляет вероятность совпадения наблюдаемых (фактических) значений и теоретических (гипотетических) значений. Если вычисленная вероятность ниже уровня значимости $(0,05)$, то нулевая гипотеза отвергается и утверждается, что наблюдаемые значения не соответствуют нормальному закону распределения. Если вычисленная вероятность близка к 1, то можно говорить о высокой степени соответствия экспериментальных данных нормальному закону распределения.

Функция имеет следующий синтаксис: **ХИ2ТЕСТ (фактический интервал; ожидаемый интервал)**

Здесь:

- **фактический интервал** — это интервал данных, которые содержат наблюдения, подлежащие сравнению с ожидаемыми значениями;

- **ожидаемый интервал** — это интервал данных, который содержит теоретические (ожидаемые) значения для соответствующих наблюдаемых.

Пример 10. (См. пример 4) Проверить соответствие выборочных данных $(64, 57, 63, 62, 58,\ 61, 63, 60, 60, 61,\ 65, 62, 62, 60, 64, 61,\ 59, 59, 63, 61, 62,\ 58, 58, 63, 61, 59, 62, 60, 60, 58, 61, 60, 63, 63, 58, 60, 59, 60, 59, 61, 62, 62, 63, 57, 61, 58, 60, 64, 60, 59, 61, 64, 62, 59, 65)$ нормальному закону распределения.

Решение.

1. Заполним следующую таблицу:

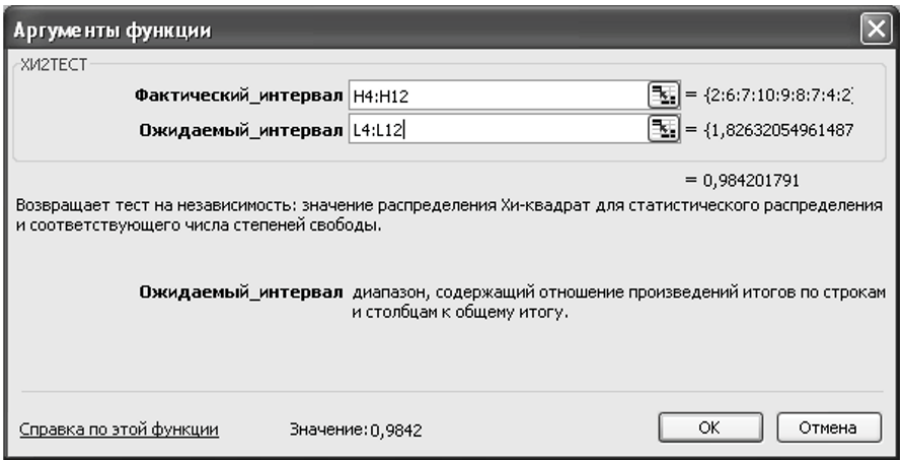
	A	B	C	D	E	F	G	H	I	J
1	**Наблюдения**						Вес	Абсолютные	Относительные	Накопленные
2	64	62	58	63	61		кг	частоты	частоты	частоты
3	57	62	63	58	58					
4	63	60	61	60	60		57	2	0,036	0,036
5	62	64	59	59	64		58	6	0,109	0,145
6	58	61	62	60	60		59	7	0,127	0,273
7	61	59	60	59	59		60	10	0,182	0,455
8	63	59	60	61	61		61	9	0,164	0,618
9	60	63	58	62	64		62	8	0,145	0,764
10	60	61	61	62	62		63	7	0,127	0,891
11	61	62	60	63	59		64	4	0,073	0,964
12	65	58	63	57	65		65	2	0,036	1,000
13								55		

2. Найдем теоретические частости нормального распределения. Для этого предварительно необходимо найти среднее значение и стандартное отклонение выборки.

В ячейке I13 с помощью функции **СРЗНАЧ** найдем среднее значение для данных из диапазона A2:E12 (60, 855). В ячейке J13 с помощью функции **СТАНДОТКЛОН** найдем стандартное отклонение для этих же данных (2, 05). В ячейки K1 и K2 введем название столбца — **Теоретические частости**. Затем с помощью функции **НОРМРАСП** найдем теоретические частости. Установим курсор в ячейку K4, вызовем указанную функцию и заполним ее рабочие поля: **x** – G4; **Среднее** — I 13; **Стандартное откл** — $J13$; **Интегральный** — 0. Получим в ячейке K4 0, 033. Далее протягиванием скопируем содержимое ячейки K4 в диапазон ячеек K5:K12. Затем в ячейки L1 и L2 введем название нового столбца – **Теоретические частоты**. Установим курсор в ячейку L4 и введем формулу =H 13*K4. Далее протягиванием скопируем содержимое ячейки L4 в диапазон ячеек L5:L12.

G	H	I	J	K	L
Вес кг	Абсолютные частоты	Относительные частоты	Накопленные частоты	Теоретические частости	Теоретические частоты
57	2	0,036	0,036	0,033205828	1,82632055
58	6	0,109	0,145	0,073795567	4,058756212
59	7	0,127	0,273	0,129258576	7,109221655
60	10	0,182	0,455	0,178443849	9,814411704
61	9	0,164	0,618	0,194158732	10,67873029
62	8	0,145	0,764	0,16650428	9,157735407
63	7	0,127	0,891	0,112540024	6,189701326
64	4	0,073	0,964	0,059951732	3,297345259
65	2	0,036	1,000	0,025171529	1,384434082
	55	60,855	2,050		

3. С помощью функции **ХИ2ТЕСТ** определим соответствие данных нормальному закону распределения. Для этого установим курсор в свободную ячейку L13 и введем функцию **ХИ2ТЕСТ**. В качестве фактического интервала зададим диапазон H4:H12, а ожидаемого интервала – диапазон L4:L12. В ячейке L13 появится значение вероятности того, что выборочные данные соответствуют нормальному закону распределения — 0, 9842.



4. Поскольку полученная вероятность соответствия экспериментальных данных $p = 0, 98$ много больше, чем уровень значимости $\alpha = 0, 05$, то можно утверждать,

что нулевая гипотеза не может быть отвергнута и, следовательно, данные не противоречат нормальному закону распределения. Более того, поскольку полученная вероятность $p = 0,98$ близка к 1, можно говорить о высокой степени вероятности того, что экспериментальные данные соответствуют нормальному закону.

3.5 Анализ двух выборок – t-критерий Стьюдента (критерий различия)

Следующей задачей статистического анализа, решаемой после определения основных выборочных характеристик и анализа одной выборки, является совместный анализ нескольких выборок. Важнейшим вопросом, возникающим при анализе двух выборок, является вопрос о наличии различий между этими выборками. Обычно для этого проводят проверку статистических гипотез о принадлежности обеих выборок одной генеральной совокупности или о равенстве генеральных средних. В рассмотренном ранее примере 6 такие различия выявляются путем сравнения данных реализации турфирмой путевок за периоды до и после начала активной рекламной кампании. Если сопоставить средние значения числа реализованных за месяц путевок до $(125,6)$ и после $(145,7)$ начала рекламной кампании, видно, что они различаются. Можно ли по этим данным сделать вывод об эффективности рекламной кампании?

Для решения задач такого типа используются так называемые критерии различия. Для проверки одной и той же гипотезы могут быть использованы разные статистические критерии. Правильный выбор критерия определяется как спецификой данных и проверяемых гипотез, так и уровнем статистической подготовки исследователя. Статистические критерии различия подразделяются на параметрические и непараметрические критерии. Параметрические критерии служат для проверки гипотез о параметрах определенных распределений генеральной совокупности (чаще всего нормального распределения). Непараметрические критерии для проверки гипотез не используют предположений о законе распределения генеральной совокупности и не требуют знания параметров распределения.

Параметрические критерии служат для проверки гипотез о положении и рассеивании. Из параметрических критериев наибольшей популярностью при проверке гипотез о равенстве генеральных средних (математических ожиданий) пользуется t-критерий Стьюдента (t-критерий различия). Он наиболее часто используется для проверки следующей гипотезы: «Средние двух выборок относятся к одной и той же совокупности». Критерий позволяет найти вероятность того, что оба средних относятся к одной и той же совокупности. Если эта вероятность p ниже уровня значимости ($p < 0,05$), то принято считать, что выборки относятся к двум разным совокупностям.

При использовании t-критерия можно выделить два случая. В первом случае его применяют для проверки гипотезы о равенстве генеральных средних двух независимых, несвязанных выборок (так называемый двухвыборочный t-критерий). В этом случае есть контрольная группа и опытная группа, состоящие, например, из разных пациентов, количество которых в группах может быть различно.

Во втором случае, когда одна и та же группа объектов порождает числовой материал для проверки гипотез о средних, используется так называемый парный t-критерий. Выборки при этом называют зависимыми, связанными. Например, из-

меряется содержание лейкоцитов у здоровых животных, а затем у тех же самых животных после облучения определенной дозой излучения.

В обоих случаях в принципе должно выполняться требование нормальности распределения исследуемого признака в каждой из сравниваемых групп и равенства дисперсий в сравниваемых совокупностях. Однако на практике по большому счету корректное применение t-критерия Стьюдента для двух групп часто бывает затруднительно, поскольку достоверно проверить эти условия удается далеко не всегда.

Для оценки достоверности отличий по критерию Стьюдента принимается нулевая гипотеза, что средние выборок равны между собой. Затем вычисляется значение вероятности того, что изучаемые события (например, количества реализованных путевок в обеих выборках) произошли случайным образом.

3.6 Использование t-критерия Стьюдента в Excel

В MS Excel для оценки достоверности отличий по критерию Стьюдента используются специальная функция **ТТЕСТ** и процедуры **Пакета анализа**. Эти перечисленные инструменты вычисляют вероятность, соответствующую критерию Стьюдента, и используются, чтобы определить, насколько вероятно, что две выборки взяты из генеральных совокупностей, которые имеют одно и то же среднее.

Функция **ТТЕСТ** имеет следующий синтаксис: **ТТЕСТ(массив1; массив2; хвосты; тип)**

Здесь:

- **массив1** — это первое множество данных;
- **массив2** — это второе множество данных;
- **хвосты** — число хвостов распределения. Обычно число хвостов равно 2;
- **тип** — это вид исполняемого t-теста. Возможны 3 варианта выбора: 1 – парный тест, 2 – двухвыборочный тест с равными дисперсиями, 3 – двухвыборочный тест с неравными дисперсиями.

Пример 11. (См. пример 6) Выявить, достоверны ли отличия при сравнении данных реализации турфирмой путевок за периоды до и после начала активной рекламной кампании.

Решение.

1. Введите данные так, как показано в следующей таблице.

	A	B
1	С рекламой	Без рекламы
2	162	135
3	156	126
4	144	115
5	137	140
6	125	121
7	145	112
8	151	130

2. Для выявления достоверности отличий установим курсор в свободную ячейку (например, А11). Вызовем **Мастер функций**, выберем категорию **Статистические**

и функцию **ТТЕСТ**. В появившемся диалоговом окне функции **ТТЕСТ** введем исходные данные: в поле **Массив1** введем диапазон A2:A8; в поле **Массив2** – диапазон данных исследуемой группы B2:B8. В поле **Хвосты** всегда вводится с клавиатуры цифра 2 (без кавычек), а в поле **Тип** с клавиатуры введем цифру 3. Нажмем кнопку **ОК**. В ячейке A11 появится значение вероятности – $0,006295$.

3. Поскольку величина вероятности случайного появления анализируемых выборок (0,006295) меньше уровня значимости ($\alpha = 0,05$), то нулевая гипотеза отвергается. Следовательно, различия между выборками не случайные и средние выборок считаются достоверно отличающимися друг от друга. Поэтому на основании применения критерия Стьюдента можно сделать вывод о большей эффективности реализации путевок после начала рекламной кампании ($p < 0,05$).

Как указывалось выше, при использовании t-критерия выделяют два основных случая. В первом случае его применяют для проверки гипотезы о равенстве генеральных средних двух независимых, несвязанных выборок (так называемый двухвыборочный t-критерий). В этом случае есть две различные выборки, количество элементов в которых может быть также различно. При заполнении диалогового окна **ТТЕСТ** при этом указывается **Тип**, равный 3.

Во втором случае, когда одна и та же группа объектов порождает числовой материал для проверки гипотез о средних, используется так называемый парный t-критерий. Выборки при этом называют зависимыми, связанными (при заполнении диалогового окна **ТТЕСТ** указывается **Тип**, равный 1. Например, сравнивается реализация путевок двумя фирмами в соответствующие месяцы.

Пример 12. Сравнивается количество наличных денег у двух групп студентов (в тыс. рублей):

30	10
30	20
40	30
50	40
60	50

.

Необходимо определить достоверность различия между группами при двух вариантах постановки задачи:

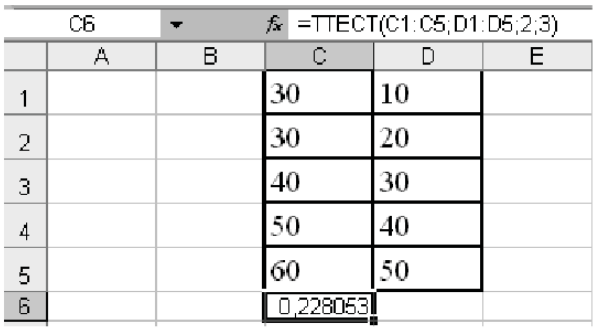
- группы состоят из различных студентов (тип 3 – двухвыборочный тест с неравными дисперсиями);

- группы состоят из одних и тех же студентов, но первая – до посещения буфета, а вторая – после (тип 1 – парный тест).

Решение.

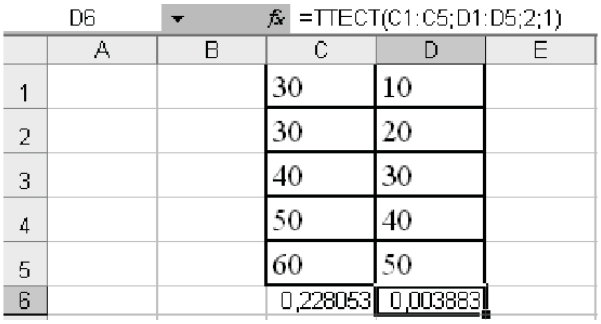
В ячейки C1:C5 введите количество денег у студентов первой группы. В ячейки D1:D5 введите количество денег у студентов второй группы.

1. Установим курсор в свободную ячейку (например, C6). Вызовем **Мастер функций**, выберем категорию **Статистические** и функцию **ТТЕСТ**. В появившемся диалоговом окне функции **ТТЕСТ** введем исходные данные. Указателем мыши введем диапазон данных первой группы в поле **Массив1** (C1:C5). В поле **Массив2** введем диапазон данных второй группы (D1:D5). В поле **Хвосты** всегда вводится цифра 2 (без кавычек), а в поле **Тип** введем цифру 3. Нажмем кнопку **ОК**. В ячейке C6 появится значение вероятности – $0,228053$.



Поскольку величина вероятности случайного появления анализируемых выборок $(0, 228053)$ больше уровня значимости $(\alpha = 0, 05)$, то нулевая гипотеза не может быть отвергнута (принимается). Следовательно, различия между выборками могут быть случайными и средние выборок не считаются достоверно отличающимися друг от друга. Поэтому на основании применения критерия Стьюдента нельзя сделать вывод о достоверности отличий двух групп студентов по количеству карманных денег, имеющихся у них $(p > 0, 05)$.

2. Установим курсор в свободную ячейку (например, D6). Вызовем **Мастер функций**, выберем категорию **Статистические** и функцию **ТТЕСТ**. В появившемся диалоговом окне функции **ТТЕСТ** введем исходные данные. Указателем мыши введем диапазон данных первой группы в поле **Массив1** (C1:C5). В поле **Массив2** введем диапазон данных второй группы (D1:D5). В поле **Хвосты** всегда вводится цифра 2, а в поле **Тип** введем цифру 1. Нажмем кнопку **ОК**. В ячейке D6 появится значение вероятности – $0, 003883$.



Поскольку величина вероятности случайного появления анализируемых выборок $(0, 003883)$ меньше уровня значимости $(\alpha = 0, 05)$, то нулевая гипотеза отвергается. Следовательно, различия между выборками не могут быть случайными и средние выборок считаются достоверно отличающимися друг от друга. Поэтому на основании применения критерия Стьюдента можно сделать вывод о том, что в двух группах студентов выявлены достоверные отличия по количеству карманных денег $(p < 0, 05)$, что явилось результатом посещения буфета.

Таким образом, ясно, что применение различных типов критерия Стьюдента может приводить к различным результатам на основании одних и тех же исходных данных. Можно предложить следующий приблизительный способ выбора типа критерия: если не ясно, какой тип критерия выбирать, выбирается тип 3; если очевидно, что выборки зависимы, связаны (например, это одни и те же студенты), то следует выбирать тип 1.

3.7 Анализ двух выборок – критерий согласия хи-квадрат

Бывают ситуации, когда необходимо сравнить две относительные или выраженные в процентах величины (доли). Примером может служить случай проверки успешности трудоустройства молодых специалистов, когда известен процент трудоустроившихся выпускников двух институтов. Для проверки достоверности различий здесь критерий Стьюдента применить не удастся. В таких задачах обычно используют критерий χ^2 (хи-квадрат).

Здесь, как и в случае с критерием Стьюдента, принимается нулевая гипотеза о том, что выборки принадлежат к одной генеральной совокупности. Кроме того, определяется ожидаемое значение результата. Обычно это среднее значение между выборками рассматриваемого показателя. Затем оценивается вероятность того, что ожидаемые значения и наблюдаемые принадлежат к одной генеральной совокупности.

В MS Excel критерий хи-квадрат реализован в функции **ХИ2ТЕСТ**. Функция **ХИ2ТЕСТ** вычисляет вероятность совпадения наблюдаемых (фактических) значений и теоретических (гипотетических) значений. Если вычисленная вероятность ниже уровня значимости $(0,05)$, то нулевая гипотеза отвергается и утверждается, что наблюдаемые значения не соответствуют теоретическим (ожидаемым) значениям.

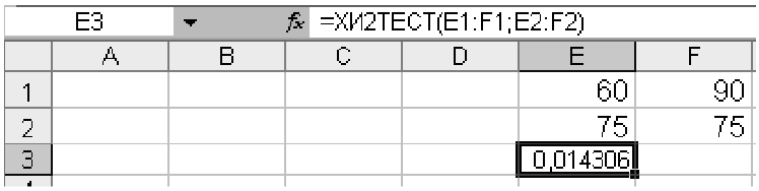
Пример 13. Пусть после окончания двух институтов экономического профиля трудоустроилось по специальности из первого института 90 человек, а из второго – 60 (обе группы молодых специалистов включали по 100 человек). Случайны ли различия между выборками.

Решение.

1. Принимается нулевая гипотеза, что выборки принадлежат к одной генеральной совокупности.

2. Определяется ожидаемое значение результата (среднее значение между выборками): $(60 + 90)/2 = 75$, то есть мы ожидали, что разницы между группами нет и в обоих случаях должно было трудоустроиться по 75 человек.

3. Затем вычисляется значение вероятности того, что изучаемые события (трудоустройство в обеих выборках) произошли случайным образом. Для этого введем данные в рабочую таблицу: 60 - в ячейку E1, 90 – в F1, 75 – в E2, F2. Установим курсор в свободную ячейку (например, E3). Вызовем **Мастер функций**, выберем категорию **Статистические** и функцию **ХИ2ТЕСТ**. В появившемся диалоговом окне функции введем исходные данные. Указателем мыши введем в поле **Фактический интервал** диапазон данных наблюдавшегося количества трудоустроившихся (E1:F1). В поле **Ожидаемый интервал** введем диапазон данных предполагаемого количества трудоустроившихся (E2:F2). Нажмем кнопку **ОК**. В ячейке E3 появится значение вероятности – $0,014306$.



Поскольку величина вероятности случайного появления анализируемых выборок $(0,0143)$ меньше уровня значимости $(\alpha = 0,05)$, то нулевая гипотеза отвергается. Следовательно, различия между выборками не могут быть случайными и выборки считаются достоверно отличающимися друг от друга. Поэтому на основании применения критерия хи-квадрат можно сделать вывод о том, что в двух группах выпускников выявлены достоверные отличия по успешности трудоустройства $(p < 0,05)$, что, по-видимому, явилось результатом более высокой репутации выпускников первого института.

Пример 14. (См. пример 7) Рассматривается заработная плата обслуживающего персонала и работников ресторана гостиницы.

Персонал	**Ресторан**
2100	3200
2100	3000
2000	2500
2000	2000
2000	1900
1900	1800
1800	
1800	

.

Можно ли по этим данным сделать вывод о большей зарплате работников ресторана?

Решение.

Для решения задач такого типа используются так называемые критерии различия, в частности, t-критерий Стьюдента.

1. Введите данные: для персонала – в диапазон A1:A8; для работников ресторана – в диапазон B1:B6.

2. Выбор процедуры осуществляется из трех вариантов t-теста. Поскольку данные не имеют попарного соответствия, число их различно и говорить о равенстве дисперсий затруднительно, выберите процедуру **Двухвыборочный t-тест с различными дисперсиями**.

Для реализации процедуры в пункте меню **Сервис** выберите строку **Анализ данных** и далее укажите курсором мыши на строку **Двухвыборочный t-тест с различными дисперсиями**.

3. В появившемся диалоговом окне задайте **Интервал переменной 1**, указывая диапазон A1:A8.

4. Аналогично укажите **Интервал переменной 2**, то есть введите ссылку на диапазон второго столбца B1:B6.

5. Далее укажите выходной диапазон. Для этого поставьте переключатель в положение **Выходной диапазон** и введите в качестве выходного диапазона ссылку на ячейку C1. Щелкните по кнопке **OK**.

Результаты анализа. В выходном диапазоне C1:E13 появятся результаты процедуры **Двухвыборочный t-тест с различными дисперсиями**.

	A	B	C	D	E
1	2100	3200	Двухвыборочный t-тест с различными дисперсиями		
2	2100	3000			
3	2000	2500		*Переменная 1*	*Переменная 2*
4	2000	2000	Среднее	1962,5	2400
5	2000	1900	Дисперсия	14107,14286	356000
6	1900	1800	Наблюдения	8	6
7	1800		Гипотетическая разность средних	0	
8	1800		df	5	
9			t-статистика	-1,769982969	
10			P(T<=t) одностороннее	0,068475305	
11			t критическое одностороннее	2,015048372	
12			P(T<=t) двухстороннее	0,13695061	
13			t критическое двухстороннее	2,570581835	

Интерпретация результатов.

Средние значения заработной платы (1962 руб. для персонала и 2400 руб. для работников ресторана) довольно сильно отличаются. Тем не менее нулевая гипотеза о том, что разницы между группами нет (то есть средние выборок равны между собой), отвергнута быть не может. Это следует из того, что вероятность реализации нулевой гипотезы достаточно велика ($p = 0,1389$, что больше чем уровень значимости $0,05$, то есть $p > 0,05$) и величина вероятности случайного появления анализируемых выборок ($P(T \leq t)$ двухстороннее) больше уровня значимости ($\alpha = 0,05$). А это позволяет говорить, что различия между выборками могут быть случайными, то есть различия недостоверные. Таким образом, из полученных результатов исследования вытекает, что на основании приведенных данных нельзя сделать вывод о достоверно большей зарплате работников ресторана.

3.8 Задания для самостоятельной работы

1. Определить, лежит ли значение 19 внутри границ 95-процентного доверительного интервала выборки $2, 3, 5, 7, 4, 9, 6, 4, 9, 10, 4, 7, 19$.

2. Определите с уровнем значимости $\alpha = 0,05$ максимальное отклонение среднего значения генеральной совокупности от среднего выборки: $3, 4, 4, 2, 5, 3, 4, 3, 5, 4, 3, 5, 6$.

3. Найти соответствие экспериментальных данных нормальному закону распределения для следующей выборки весов детей (кг):

$21, 21, 22, 22, 22,\ 22, 22, 22, 22, 22, 22,\ 23, 23, 23, 23, 23,\ 23, 24, 24, 24, 24,\ 24, 24, 24, 24,$
$24, 24, 24, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 25, 26, 26, 26, 26, 26, 26, 26,$
$26, 26,\ 26, 26, 26,\ 26, 26, 27, 27.$

4. Даны результаты бега на дистанцию 100 м в секундах в двух группах студентов. Студенты первой группы в течение года посещали факультативные занятия по физкультуре. Определить, достоверны ли отличия по результатам бега в этих группах.

Посещавшие факультатив	Не посещавшие факультатив
12,6	12,8
12,3	13,2
11,9	13,0
12,2	12,9
13,0	13,5
12,4	13,1

.

5. В ходе социологического опроса на вопрос о перенесенном в детстве заболевании ответы распределились следующим образом:

	Да	Нет	Не помню
Мужчины	58	11	10
Женщины	35	25	23

.

Есть ли достоверные отличия в ответах женщин и мужчин?

6. Приведены данные ежемесячной результативности (количество голов) футбольной команды в двух сезонах

Месяц	3	4	5	6	7	8	9	10	11
2008 г.	3	4	5	8	9	1	2	4	5
2009 г.	6	19	3	2	14	4	5	17	1

.

Определить, есть ли статистические различия в ежемесячной результативности команды в рассматриваемых сезонах?

7. Определить, достоверны ли различия в количестве приобретаемых туристских путевок семейными парами и отдельными туристами.

	Количество приобретаемых путевок					
Месяцы	**1**	**2**	**3**	**4**	**5**	**6**
Пары	67	75	58	89	96	94
Одиночки	43	56	78	87	85	90

.

8. В таблице приведены результаты группы студентов по скоростному чтению до и после специального курса по быстрому чтению.

Студент	1	2	3	4	5	6	7	8	9	10
До курса	86	83	86	70	66	90	70	85	77	86
После курса	82	79	91	77	68	86	81	90	85	94

.

Произошли ли статистически значимые изменения скорости чтения у студентов?

4 Дисперсионный анализ

В случае необходимости оценить достоверность различия между несколькими группами наблюдений (выборками) используют методы дисперсионного анализа.

Дисперсионный анализ предназначен для исследования задачи о действии на измеряемую случайную величину (отклик) одного или нескольких независимых факторов, имеющих несколько градаций. Причем в однофакторном, двухфакторном и т. д. анализе влияющие на результат факторы считаются известными и речь идет только о выяснении существенности или оценке этого влияния.

Применение дисперсионного анализа возможно, если можно предполагать соответствие выборочных групп генеральным совокупностям с нормальным распределением и независимость распределений наблюдений в группах.

В дальнейшем ограничимся рассмотрением простейшего случая дисперсионного анализа - однофакторного анализа. При этом задача заключается в том, чтобы сравнить дисперсию, обусловленную случайными причинами, с дисперсией, вызываемой наличием исследуемого фактора. Если они значимо различаются, то считают, что фактор оказывает статистически значимое влияние на исследуемую переменную. Значимость различий проверяется по критерию Фишера.

Критерий Фишера используют для проверки гипотезы о принадлежности двух дисперсий одной генеральной совокупности и, следовательно, их равенстве. При этом предполагается, что данные независимы и распределены по нормальному закону. Гипотеза о равенстве дисперсий принимается, если отношение большей дисперсии к меньшей меньше критического значения распределения Фишера:

$$F = s_1^2/s_2^2,\ F < F_{\text{крит}},$$

где $F_{\text{крит}}$ зависит от уровня значимости и числа степеней свободы для дисперсий в числителе и знаменателе.

В MS Excel для расчета уровня вероятности выполнения гипотезы о равенстве дисперсий могут быть использованы функция **ФТЕСТ(массив1; массив2)** и процедура **Пакета анализа Двухвыборочный F-тест для дисперсий**.

Пример 15. Необходимо выявить, влияет ли расстояние от центра города на степень заполняемости гостиниц. Пусть введены 3 уровня расстояний от центра города: 1) до 3 км, 2) от 3 до 5 км и 3) свыше 5 км. Данные заполняемости представлены в таблице.

Расстояние	**Заполняемость, %**					
до 3 км	92	98	89	97	90	94
от 3 до 5 км	90	86	84	91	83	82
свыше 5 км	97	79	74	85	73	77

.

Решение.

1. Исследуемые данные введите в рабочую таблицу Excel по столбцам: в столбец А – заполняемость гостиниц в центре города, в столбец В – гостиниц, находящихся на расстоянии от 3 до 5 км и т. д. (диапазон A1:C6).

2. Выполните команду **Сервис - Анализ данных**. В появившемся диалоговом окне **Анализ данных** в списке **Инструменты анализа** щелчком мыши выберите процедуру **Однофакторный дисперсионный анализ**. Нажмите кнопку **ОК**.

3. В появившемся диалоговом окне **Однофакторный дисперсионный анализ** в поле **Входной интервал** задайте A1:C6.

4. В разделе **Группировка** переключатель установите в положение **по столбцам**.

5. Далее необходимо указать выходной диапазон. Для этого поставьте переключатель в положение **Выходной интервал**, затем щелкните указателем мыши в правом поле ввода **Выходной интервал** и щелчком мыши на ячейке A8 укажите расположение выходного диапазона. Нажмите кнопку **OK**.

	A	B	C	D	E	F	G
1	92	90	87				
2	98	86	79				
3	89	84	74				
4	97	91	85				
5	90	83	73				
6	94	82	77				
7							
8	Однофакторный дисперсионный анализ						
9							
10	ИТОГИ						
11	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*		
12	Столбец 1	6	560	93,33333	13,46666667		
13	Столбец 2	6	516	86	14		
14	Столбец 3	6	475	79,16667	32,96666667		
15							
16							
17	Дисперсионный анализ						
18	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-Значение*	*F критическое*
19	Между группами	602,3333	2	301,1667	14,95035852	0,000268401	3,682320344
20	Внутри групп	302,1667	15	20,14444			
21							
22	Итого	904,5	17				

Интерпретация результатов. В таблице "Дисперсионный анализ" на пересечении строки "Между группами" и столбца "P-Значение"находится величина $0,0002684$. Величина P-Значение $< 0{,}05$, следовательно, критерий Фишера значим и влияние фактора расстояния от центра города на эффективность заполнения гостиниц доказано статистически.

5 Корреляционный анализ

Одна из наиболее распространенных задач статистического исследования состоит в изучении связи между некоторыми наблюдаемыми переменными. Знание взаимозависимостей отдельных признаков дает возможность решать одну из кардинальных задач любого научного исследования: возможность предвидеть, прогнозировать развитие ситуации при изменении конкретных характеристик объекта исследования. Например, основное содержание любой экономической политики в конечном счете может быть сведено к регулированию экономических переменных, осуществляемому на базе выявленной тем или иным образом информации об их взаимовлиянии. Поэтому проблема изучения взаимосвязей показателей различного рода является одной из важнейших в статистическом анализе.

Обычно взаимосвязь между выборками носит не функциональный, а вероятностный (или стохастический) характер. В этом случае нет строгой, однозначной зависимости между величинами. При изучении стохастических зависимостей различают регрессию и корреляцию.

Регрессионный анализ устанавливает формы зависимости между случайной величиной Y и значениями одной или нескольких переменных величин.

Корреляционный анализ состоит в определении степени связи между двумя случайными величинами X и Y. В качестве меры такой связи используется коэффициент корреляции. Он оценивается по выборке объема n связанных пар наблюдений (x_i, y_i) из совместной генеральной совокупности X и Y. Существует несколько типов коэффициентов корреляции, применение которых зависит от предположений о совместном распределении величин X и Y.

Для оценки степени взаимосвязи наибольшее распространение получил коэффициент линейной корреляции (Пирсона), предполагающий нормальный закон распределения наблюдений.

Коэффициент корреляции r – параметр, характеризующий степень линейной взаимосвязи между двумя выборками. Коэффициент корреляции изменяется от -1 (обратная зависимость) до 1 (прямая зависимость). При значении коэффициента равном 0 линейной зависимости между двумя выборками нет. Здесь под прямой зависимостью понимают зависимость, при которой увеличение или уменьшение значения одного признака ведет, соответственно, к увеличению или уменьшению второго. Например, при увеличении температуры возрастает давление газа, а при уменьшении – снижается. При обратной зависимости увеличение одного признака приводит к уменьшению второго и наоборот. Примером обратной корреляционной зависимости может служить связь между температурой воздуха на улице и количеством топлива, расходуемого на обогрев помещения.

Коэффициент корреляции является безразмерной величиной, и его значение не зависит от единиц измерения случайных величин X и Y.

Для оценки степени взаимосвязи можно руководствоваться следующими эмпирическими правилами. Если коэффициент корреляции r по абсолютной величине (без учета знака) больше, чем $0,95$, то принято считать, что между параметрами существует практически линейная зависимость (прямая - при положительном r и обратная - при отрицательном r). Если абсолютная величина коэффициента корреляции $|r|$ лежит в диапазоне от $0,8$ до $0,95$, говорят о сильной степени линейной связи между параметрами. Если $0,6 < |r| < 0,8$, говорят о наличии линейной связи между параметрами. При $|r| < 0,4$ обычно считают, что линейную взаимосвязь между параметрами выявить не удалось.

В MS Excel для вычисления парных коэффициентов линейной корреляции используется специальная функция **КОРРЕЛ**.

Функция имеет следующий синтаксис: **КОРРЕЛ(массив1; массив2)**

Здесь:

- **массив1** – это диапазон ячеек первой случайной величины;

- **массив2** – это второй интервал ячеек со значениями второй случайной величины.

Пример 16. Имеются результаты семимесячных наблюдений реализации путевок двух туристских маршрутов тура А и тура В, представленные в следующей таблице:

Тур А	120	121	105	92	112	91	80
Тур В	20	19	17	16	18	16	15

.

Необходимо определить, имеется ли взаимосвязь между количеством продаж путевок обоих маршрутов.

Решение.

1. Для выявления степени взаимосвязи прежде всего необходимо ввести данные в рабочую таблицу.

	A	B
1	**Тур А**	**Тур В**
2	120	20
3	121	19
4	105	17
5	92	16
6	112	18
7	91	16
8	80	15

2. Затем вычисляется значение коэффициента корреляции между выборками. Для этого установите курсор в свободную ячейку (например, А9). Вызовите функцию **КОРРЕЛ**. Введите в поле **Массив1** диапазон данных А2:А8. В поле **Массив2** введите диапазон данных В2:В8. Нажмите кнопку **ОК**.

3. В ячейке А9 появится значение коэффициента корреляции – $0,969123$. Значение коэффициента корреляции больше чем $0,95$. Значит, можно говорить о том, что в течение периода наблюдения имелась высокая степень прямой линейной взаимосвязи между количествами проданных путевок обоих маршрутов ($r = 0,969123$).

Пример 17. Имеются ежемесячные данные наблюдений за состоянием погоды и посещаемостью музеев и парков.

Число ясных дней	8	14	20	25	20	15
Количество посетителей музея	495	503	380	305	348	465
Количество посетителей парка	132	348	643	865	743	541

.

Необходимо определить, существует ли взаимосвязь между состоянием погоды и посещаемостью музеев и парков.

Решение.

1. Для выполнения корреляционного анализа введите в диапазон А1:G3 исходные данные:

	A	B	C	D	E	F	G
1	**Число ясных дней**	**8**	**14**	**20**	**25**	**20**	**15**
2	**Количество посетителей музея**	**495**	**503**	**380**	**305**	**348**	**465**
3	**Количество посетителей парка**	**132**	**348**	**643**	**865**	**743**	**541**

2. Затем выполните команду **Сервис - Анализ данных** и выберите строку **Корреляция**. В появившемся диалоговом окне укажите **Входной интервал** В1:G3. Укажите, что данные рассматриваются **по строкам**. Укажите выходной диапазон. Для этого поставьте флажок в левое поле **Выходной интервал** и в правое поле ввода **Выходной интервал** введите А4. Нажмите кнопку **ОК**.

4		*Строка 1*	*Строка 2*	*Строка 3*
5	Строка 1	1		
6	Строка 2	-0,92185	1	
7	Строка 3	0,974576	-0,91938	1

Интерпретация результатов. Из таблицы видно, что корреляция между состоянием погоды и посещаемостью музея равна $-0,92$, а между состоянием погоды и посещаемостью парка – $0,97$, между посещаемостью парка и музея – $-0,92$. Таким образом, в результате анализа выявлены зависимости: сильная степень обратной линейной взаимосвязи между посещаемостью музея и количеством солнечных дней ($r = -0,92$) и практически линейная (очень сильная прямая) связь между посещаемостью парка и состоянием погоды ($r = 0,97$). Между посещаемостью музея и парка имеется сильная обратная взаимосвязь ($r = -0,92$). Подразумевается, что в пустых клетках в правой верхней половине таблицы находятся те же коэффициенты корреляции, что и в нижней левой (симметрично расположенные относительно диагонали).

5.1 Задания для самостоятельной работы

1. Определить, влияет ли фактор образования на уровень зарплаты сотрудников в гостинице на основании следующих данных:

Образование	**Зарплата сотрудника**					
высшее	32000	30000	26000	20000	19000	19000
ср. проф.	26000	20000	20000	19000	18000	17000
среднее	20000	20000	19000	18000	17000	17000

.

2. Определить, имеется ли взаимосвязь между рождаемостью и смертностью (количество на 1000 человек) в Санкт-Петербурге:

Годы	**Рождаемость**	**Смертность**
2011	9,3	12,5
2012	7,4	13,5
2013	6,6	17,4
2014	7,1	17,2
2015	7,0	15,9
2016	6,6	14,2

.

3. Определить, имеется ли взаимосвязь между годовым уровнем инфляции (%), ставкой рефинансирования (%) и курсом доллара (руб./$), по следующим данным ежегодных наблюдений:

Уровень инфляции	**Ст. рефин.**	**Курс $**
84	85	42,5
45	55	52,3
56	65	56,4
34	40	60,2
23	28	65,9

.

6 Литература, Интернет-ресурсы

1. Боровков В.П. Популярное введение в программу STATISTICA. - М.: КомпьютерПресс, 1998.

2. Гарнаев А.Ю. Excel, VBA, Internet в экономике и финансах. – СПб.: БХВ-Петербург, 2005.

3. Гельман В.Я. Решение математических задач средствами Excel: практикум. – СПб.: Питер, 2003.

4. Гмурман В.Е. Теория вероятностей и математическая статистика: учеб. пособие. - М.: Юрайт, 2011.

5. Крыштановский А.О. Анализ социологических данных с помощью пакета SPSS. - М.: ИД ГУ ВШЭ, 2006.

6. Малхотра Н. Маркетинговые исследования. - М.: Вильямс, 2003.

7. Мидлтон М.Р. Анализ статистических данных с использованием Microsoft Excel для Office XP. – М.: БИНОМ. Лаборатория знаний, 2005.

8. Основы компьютерных технологий в образовании. Статистический анализ и обработка данных с применением MS Excel: учеб. пособие / С.И. Максимов [и др.]. – Минск: РИВШ БГУ, 2006.

Перечень ресурсов сети Интернет

Сайт В.С. Аванесова

Социологические исследования

Социология и маркетинг в сети

Социология: методология, методы, математические модели

Статистические методы. Сайт А.И. Орлова

StatSoft Russia