ПРЕДИСЛОВИЕ

Любому специалисту в ходе практической деятельности приходится совершать операции над количественными данными, которые осуществляются в соответствии с математическими законами. Поэтому для специалиста-нематематика наиболее важным является практический аспект математики и умение провести необходимые вычисления. Математическая теория изменяется сравнительно медленно, однако технология применения математических методов претерпела более существенные изменения. В настоящее время специалист, даже хорошо знающий математику, но не умеющий применять математические методы на компьютере, не может считаться специалистом современного уровня.

Настоящее пособие посвящено описанию методов проведения анализа экспериментальных данных и их реализации с помощью пакета **Microsoft Excel**. Наиболее важной отличительной особенностью предлагаемого в учебном пособии материала является рассмотрение основных разделов статистической обработки экспериментальных данных не в традиционном изложении, а с перспективой дальнейшего применения компьютера. При этом изложение материала ведется от математической постановки задач к способам их решения на компьютере.

Существует значительное количество специализированных математических пакетов, таких как MatLab, MathCad, Mathematica, Maple и др. Все они охватывают основные разделы математики и позволяют производить подавляющее большинство необходимых математических расчетов. Однако освоение этих пакетов самостоятельно – довольно трудоемкая задача. В то же время в курс «Информационные технологии» (или «Основы информационных технологий»), изучаемый в различных вузах страны, включено изучение прикладной программы по расчету в таблицах **Microsoft Excel**. Поэтому представляется оправданным описанный в данном пособии подход, основанный на применении математических методов именно с помощью программы **Excel**.

Настоящее пособие предназначено, в первую очередь, для студентов и магистрантов Государственного института управления и социальных технологий Белорусского государственного университета и ориентировано на дальнейшее использование информационных технологий в процессе обучения выбранной специальности («Менеджмент», «Социальная работа», «Правоведение»). Однако оно может быть рекомендовано широкому кругу пользователей персонального компьютера, сталкивающихся с необходимостью математической обработки данных

Данное пособие содержит основные теоретические сведения о таких математических методах статистического анализа, как: аппроксимация экспериментальных данных, дисперсионный, корреляционный и регрессионный анализ. Изложение материала осуществляется в следующей последовательности. Вначале приводятся основные определения и формулы, затем дается описание соответствующих процедур и функций **Microsoft Excel**, после чего рассматриваются решения типовых примеров.

В пособии предложены задания для самостоятельной работы студентов.

1. ОБЩИЕ ПОЛОЖЕНИЯ

Представьте себе, что вы забыли таблицу умножения и решили ее освежить в памяти. Но вот беда: на обложке старой тетради сохранилась лишь часть таблицы. Что вы станете делать?

Перед нами оставшаяся часть таблицы умножения на 5:

$5 \times 1 = 5$

$5 \times 2 = 10$

$5 \times 3 = 15$

$5 \times 4 = 20$ ….

Дальше таблица обрывается. Но ведь это не страшно. Даже если мы и забыли, сколько будет 5×5, все же можно, глядя на таблицу, сообразить, что каждый следующий результат будет отличаться от предыдущего на 5. Значит, после 20 должно быть 25, затем 30 и т. д.

Такой переход **от того, что было, к тому, что будет**, называется **экстраполяцией.** Мы как бы говорим: вот что получится в будущем, если и дальше все пойдет, как было прежде.

Рассмотрим еще один пример. Пусть требуется узнать, сколько людей будет жить на Земле через некоторое время, скажем, к 2015 году. Это не только интересно, но и весьма важно для экономики. Попробуем провести расчет **методом экстраполяции**. Возьмем листок миллиметровой бумаги и станем отмечать по горизонтальной оси годы, а по вертикальной – количество людей. Найдем точки пересечения каждого года с числом людей, которые в это время жили на Земле. Точки соединим плавной кривой линией. Эта кривая – график роста народонаселения нашей планеты. Однако довести кривую можно лишь до того года, когда была последняя перепись населения. Что будет дальше, никто не знает.

Вспомним правило **экстраполяции**: «дальше – как раньше» и плавно продолжим нашу кривую, сохраняя ее форму. Продолжение сделаем не сплошной линией, а пунктиром. Ведь это лишь предположение. Но и оно оказывается весьма полезным. Теперь по нашему графику мы можем узнать, сколько примерно людей будет нас окружать в будущем, в том числе и в 2015 году.

Экстраполяция способна работать далеко не всегда. Так и в нашем примере роста народонаселения на планете: в 1900 году жило 1,5 млрд чел., в 1950 году – 2,5 млрд чел., в 1960 – 3 млрд, в 1970 – 3,5 млрд, в 1976 году появился четырехмиллиардный житель Земли. При таких темпах число людей на Земле удваивается примерно каждые 35 лет. Если продолжить с помощью экстраполяции этот процесс в будущее, то получится следующее. Один видный американский ученый подсчитал, что если рост человечества и дальше будет идти такими же темпами, то

13 июля 2116 года в мире не останется места, где бы мог стоять очередной житель Земли. Это, конечно, явная чепуха. Очевидно, что помимо экстраполяции необходимо уметь учитывать какие-то более сложные закономерности роста народонаселения, закономерности, не укладывающиеся в столь простые схемы.

Экстраполяция широко применяется в экономических прогнозах будущего спроса и предложения, а также рыночной стоимости товаров, услуг, курсов ценных бумаг и т. д.

Давайте рассмотрим еще один пример. Пусть далее в качестве объекта исследования выступает ***недвижимость***. Стоимость недвижимости определяется совокупностью множества ее характеристик (факторов), которые могут быть независимыми или зависимыми друг от друга. Агент по продаже недвижимости мог бы вносить в каждый элемент своего так называемого реестра размер дома (в квадратных метрах), число спален, средний доход населения в этом районе в соответствии с данными переписи и субъективную оценку привлекательности дома. Как только эта информация собрана для различных домов, было бы интересно посмотреть, связаны ли и каким образом эти характеристики дома с ценой, по которой он был продан. Например, могло бы оказаться, что число спальных комнат является лучшим предсказывающим фактором для цены продажи дома в некотором специфическом районе, чем «привлекательность» дома (субъективная оценка). Могли бы также обнаружиться и так называемые «выбросы», т. е. дома, которые могли бы быть проданы дороже, учитывая их расположение и характеристики.

Например, удаленность жилого дома от станции метро влияет определенным образом на рыночную стоимость дома независимо от его физических характеристик, а такая характеристика, как количество комнат в квартире, зависит от ее общей площади. В первом случае влияние рассматриваемого фактора может быть выражено количественно как реакция открытого рынка в денежном выражении, во втором случае количественное выражение влияния на стоимость данного фактора из рыночных данных выделить сложно.

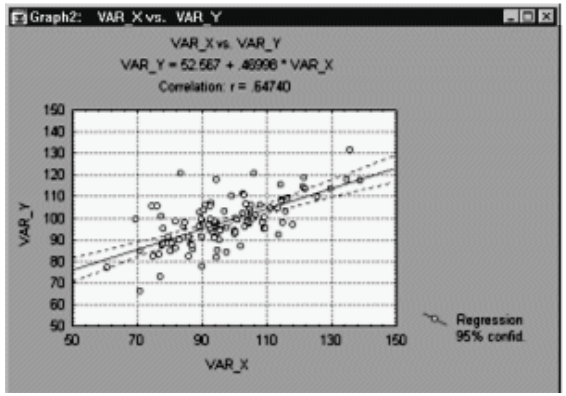
При применении, например, сравнительного подхода при осуществлении оценки основной задачей оценщика является определение стоимости выбранной единицы сравнения, которая, в свою очередь, зависит от характеристик сделок и параметров объектов сравнения. Процесс оценки предполагает для выбранных объектов сравнения выявление основных элементов сравнения, определение реакции рынка на их присутствие и корректировку выбранных единиц сравнения с учетом выявленных элементов. В итоге на последнем этапе расчетов оценщик получает диапазон значений стоимости единицы сравнения для объектов-аналогов.

Для получения достоверного результата в общем случае следует анализировать все имеющиеся продажи на данном рынке для рассматриваемого типа недвижимости. Однако на практике обычно используют выборку рыночных данных без соответствующего обоснования ее размеров, что приводит к неадекватной оценке реакции рынка на ту или иную характеристику объекта недвижимости. В конечном итоге оценщик может получить значение стоимости единицы сравнения, не

соответствующее среднерыночному значению для данного типа недвижимости на конкретном рынке.

Вместе с тем стоимость недвижимости, которая зависит от изменения множества случайных значений определяющих факторов, сама является случайной величиной, подчиняющейся действию математических законов для случайных величин. Поэтому для определения стоимости недвижимости (как случайной величины) на основе анализа рыночных данных можно применять методы математической статистики. Статистический анализ рыночных данных позволяет избежать ошибок и приблизить полученный результат к действительной реакции рынка на характеристики оцениваемого объекта недвижимости. При этом можно обоснованно получать решения на основании ограниченной выборки рыночных данных. Таким образом, при использовании методов математической статистики имеется возможность существенно расширить диапазон возможностей и качество работы оценщика при анализе данных.

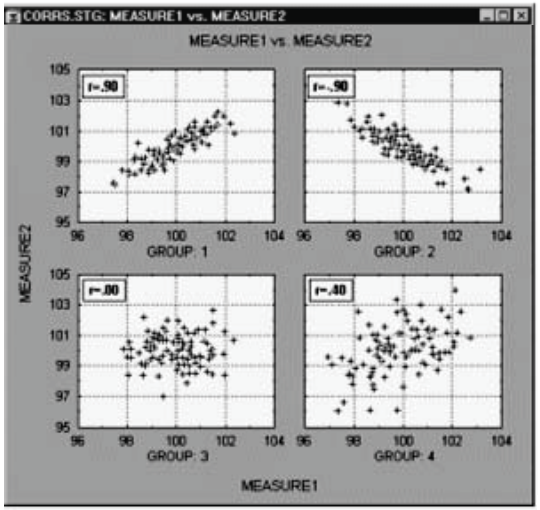
Рассмотрим применение методов статистического анализа на следующем примере. Допустим, мы имеем скорректированное на дату оценки множество значений стоимости единицы сравнения для объектов сравнения Y: $y_1, y_2, \ldots, y_n$, полученных для множества значений основного определяющего фактора X: $x_1, x_2, \ldots, x_n$, например, площади.



Две координаты, которые определяют положение каждой точки, соответствуют значениям двух переменных. Если две переменные сильно связаны, то множество точек данных принимает определенную форму (например, прямой линии или кривой). Если же переменные не связаны, то точки образуют «облако».

Если мы хотим определить стоимость единицы сравнения y_k для оцениваемого объекта, имеющего промежуточное значение определяющего фактора x_k, мы должны анализировать данные только для значений X, близких к x_k, то есть использовать только объекты сравнения, близкие к оцениваемому по основному определяющему

фактору, при этом остальная часть выборки не рассматривается. С учетом имеющихся данных (координаты точки $F(x_k,y_k)$), единственное аппроксимирующее уравнение, которое мы можем подобрать в рассматриваемом случае, является уравнением прямой $\boldsymbol{y = ax}$, проходящей через точку $F(x_k, y_k)$. Попытка вычислений по этому уравнению даже в ближайшей области x_k приводит к ошибочным результатам.



З а м е ч а н и е . **Задача аппроксимации** – это задача сглаживания экспериментальных данных. Более точно: **аппроксимацией** называется процесс подбора эмпирической формулы $\boldsymbol{\varphi(x)}$ для установленной из опыта функциональной зависимости $\boldsymbol{Y = f(x)}$. Формула служит для аналитического представления опытных данных.

Однако данную выборку можно аппроксимировать более сложными зависимостями. Например, показанные на рис. 1.1 данные можно аппроксимировать прямой $\boldsymbol{y = ax + b}$.

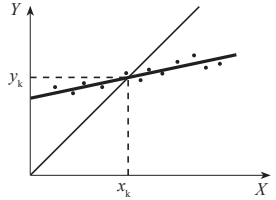


Рис. 1.1. Аппроксимация статистических данных

Тогда мы можем, используя эту прямую, определить величину стоимости объекта оценки для любого значения определяющего фактора x_i, лежащего в интервале $x_1, x_2,..., x_n$, а во многих случаях и для экстраполированных значений X. Таким образом, оценщик получает возможность использовать всю выборку значений стоимости единицы сравнения, что более полно отражает состояние рынка.

Предсказание значений по имеющимся данным осуществляется с помощью регрессионного анализа. **Регрессионный анализ** – это статистический метод, позволяющий найти уравнение, наилучшим образом описывающее множество данных.

Регрессионный анализ устанавливает формы зависимости между случайной величиной Y (зависимой) и значениями одной или нескольких переменных величин (независимых), причем значения последних считаются точно заданными. Такая зависимость обычно определяется некоторой математической моделью – ***уравнением регрессии***, содержащим несколько неизвестных параметров. В ходе регрессионного анализа на основании выборочных данных находят оценки этих параметров, определяются статистические ошибки оценок или границы доверительных интервалов и проверяется соответствие (адекватность) принятой математической модели экспериментальным данным.

В ***линейном регрессионном анализе*** связь между случайными величинами предполагается ***линейной***. В самом простом случае в линейной регрессионной модели имеются две переменные X и Y. И требуется по n парам наблюдений (X_1, Y_1), (X_2, Y_2), ... ,(X_n, Y_n) построить (подобрать) прямую линию, называемую ***линией регрессии***, которая «наилучшим образом» приближает наблюдаемые значения. Уравнение этой линии $\boldsymbol{Y = aX + b}$ является **регрессионным уравнением.** С помощью регрессионного уравнения можно предсказать ожидаемое значение зависимой величины Y_k, соответствующее заданному значению независимой переменной X_k.

Таким образом, можно сказать, что линейный регрессионный анализ заключается в подборе графика и его уравнения для набора наблюдений.

В случае, когда рассматривается зависимость между одной зависимой переменной Y и несколькими независимыми переменными $X_1, X_2, ... ,X_n$, говорят о **множественной линейной регрессии**. В этом случае регрессионное уравнение имеет вид

$$\boldsymbol{Y = a_0 + a_1X_1 + a_2X_2 +... + a_nX_n},$$

где $a_1, a_2, ... , a_n$ – требующие определения коэффициенты при независимых переменных $X_1, X_2, ... , X_n$;

a_0 – константа.

2. АППРОКСИМАЦИЯ ЭКСПЕРИМЕНТАЛЬНЫХ ДАННЫХ

На практике часто приходится сталкиваться с задачей о сглаживании экспериментальных зависимостей или задачей **аппроксимации**. Рассмотрим более подробно, что это такое, и каким образом она реализуется средствами программы табличного процессора **Microsoft Excel.**

Одна независимая переменная

Обычно задача аппроксимации распадается на две составляющие. Сначала устанавливают вид зависимости $y = f(x)$ и, соответственно, вид эмпирической формулы, то есть решают, является ли она линейной, квадратичной, логарифмической или какой-либо другой. После этого определяются численные значения неизвестных параметров выбранной эмпирической формулы, для которых приближение к заданной функции оказывается наилучшим. Если нет каких-либо теоретических соображений для подбора вида формулы, обычно выбирают функциональную зависимость из числа наиболее простых, сравнивая их графики с графиком заданной функции.

После выбора вида формулы определяют ее параметры. Для наилучшего выбора параметров задают меру близости аппроксимации экспериментальных данных. Во многих случаях, в особенности если функция $f(x)$ задана графиком или таблицей (на дискретном множестве точек), для оценки степени приближения рассматривают разности $\boldsymbol{f(x_i) - \varphi(x_i)}$ для точек $x_0, x_1, ..., x_n$. Существуют различные меры близости и, соответственно, способы решения этой задачи. Некоторые из них очень просты, быстро приводят к результату, но результат этот является сильно приближенным. Другие более точные, но и более сложные. Обычно оценку определения параметров при известном виде зависимости осуществляют по **методу наименьших квадратов**. При этом функция $\varphi(x)$ считается наилучшим приближением к $f(x)$, если для нее сумма квадратов отклонений «теоретических» значений $\varphi(x_i)$, найденных по эмпирической формуле, от соответствующих опытных значений y_i:

$$Z = \sum_{i=0}^{n} [f(x_i) - \varphi(x_i)]^2 \to \min \qquad (2.1)$$

имеет наименьшее значение по сравнению с другими функциями, из числа которых выбирается искомое приближение.

Метод наименьших квадратов формулирует аналитические условия достижения суммой квадратов отклонений (2.1) своего наименьшего значения. Так, если функция $\varphi(x)$ вполне определяется своими параметрами $k, l, m, ...,$ то наилучшие

(в указанном смысле (2.1) значения этих параметров находятся из решения системы уравнений. Например, в простейшем случае, когда функция $\varphi(x)$ представлена линейным уравнением $y = ax+b$, система имеет вид:

$$\begin{cases} a \cdot \sum_{i=1}^{n} x_i^2 + b \cdot \sum_{i=1}^{n} x_i = \sum_{i=1}^{n} x_i \cdot y_i \\ a \cdot \sum_{i=1}^{n} x_i + b \cdot n = \sum_{i=1}^{n} y_i . \end{cases} \qquad (2.2)$$

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 отражает следующая таблица:

$\boldsymbol{x}$	x_1	…	x_n
$\boldsymbol{y}$	y_1	…	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 $y_i - y_i(x_i)$ в каком-то смысле должны быть наименьшими, например в смысле (2.1).

Выяснить вид функции можно либо из теоретических соображений, либо анализируя расположение точек $(x_i; y_i)$ на координатной плоскости.

Например, пусть точки расположены так, как показано на рис. 2.1.

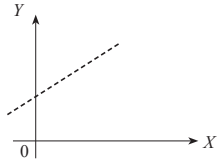


Рис. 2.1. Возможный вариант расположения экспериментальных точек

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 $\boldsymbol{y = ax + b}$.

Чтобы функция приняла конкретный вид, необходимо каким-то образом вычислить a и b. Для этого можно решить систему (2.2).

Расположение экспериментальных точек в виде кривой на рис. 2.2 наводит на мысль, что зависимость обратно пропорциональна и функцию $\varphi(x)$ нужно подбирать в виде $\boldsymbol{y = a + b/x}$. Здесь также необходимо вычислить параметры a и b.

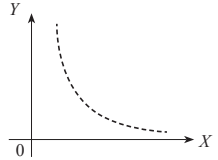


Рис. 2.2. Другой вариант расположения экспериментальных точек

Таким образом, расположение экспериментальных точек может иметь самый различный вид, и каждому соответствует конкретный тип функции.

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 ($n \le 6$), 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:

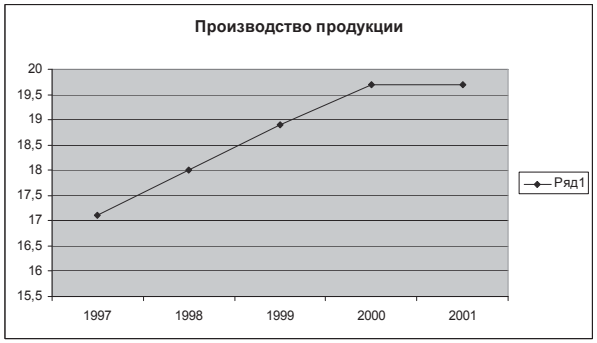
Год	Производство продукции
1997	17,1
1998	18,0
1999	18,9
2000	19,7
2001	19,7

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

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.

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. 2.3).

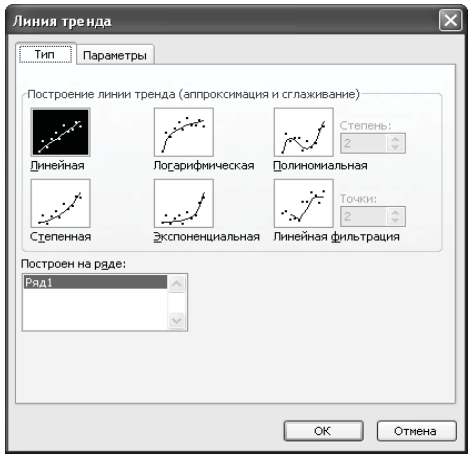


Рис. 2.3. Вкладка **Тип** диалогового окна **Линия тренда**

В этом окне на вкладке **Тип** выбираем тип линии тренда – **Полиномиальная** – и устанавливаем степень – 2. Затем открываем вкладку **Параметры** (рис. 2.4) и устанавливаем флажки в поля **показывать уравнение на диаграмме** и **поместить на диаграмму величину достоверности аппроксимации (R^2).**

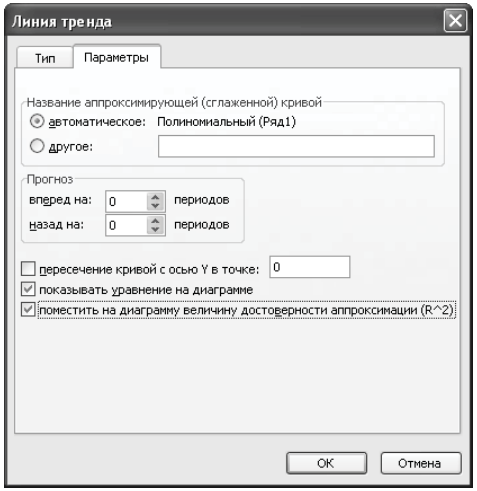


Рис. 2.4. Вкладка **Параметры** диалогового окна **Линия тренда**

После чего нужно щелкнуть по кнопке ОК. В результате получим на диаграмме аппроксимирующую кривую (рис. 2.5).

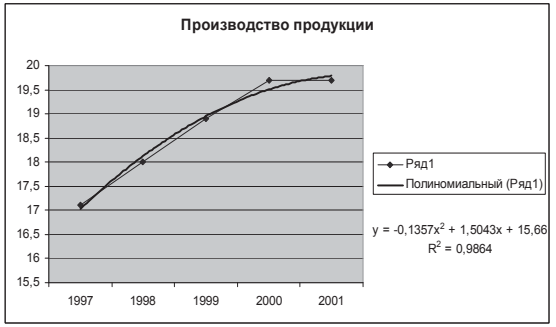


Рис. 2.5.Экспериментальные данные,
аппроксимированные полиномиальной кривой, из примера 1

Как видно из рис. 2.5, уравнение наилучшей полиномиальной аппроксимирующей функции для некоторых отвлеченных значений x (1, 2, 3, ...) выглядит как

$$y = -0{,}14x^2 + 1{,}5x + 15{,}66.$$

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. 2.6).

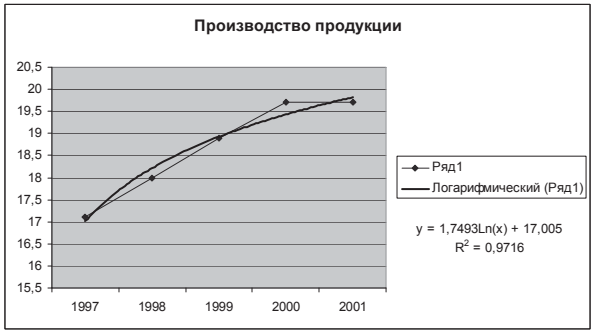


Рис. 2.6. Экспериментальные данные,
аппроксимированные логарифмической кривой, из примера 1

Как можно видеть из рис. 2.6, 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).

Несколько независимых переменных

В тех случаях, когда аппроксимируемая переменная y зависит от нескольких независимых переменных $x_1, x_2, \ldots, x_n$, т. е. $\boldsymbol{y = f(x_1, x_2, ..., x_n)}$, 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:

$$y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_nx_n, \qquad (2.3)$$

ЛГРФПРИБЛ и **РОСТ** для аппроксимации показательных функций вида:

$$y = a_0 a_1^{x_1} a_2^{x_2} \ldots a_n^{x_n}. \qquad (2.4)$$

Функции **ЛИНЕЙН** и **ЛГРФПРИБЛ** служат для вычисления неизвестных коэффициентов $a_0, \ldots, a_n$ в выражениях (2.3) и (2.4) соответственно, 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 (2.3) и (2.4) соответственно 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 соответствующие 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 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:

x_1/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 уравнением вида (2.4) 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 (рис. 2.7).

	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. 2.7. Исходные данные из примера 2

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:

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

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).

x_1	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 y 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 (рис. 2.8).

	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	3,3
11	105	15	3,3
12	110	15	3,4
13	115	15	3,4
14	120	15	3,4
15	125	15	3,5
16	130	15	3,5
17	100	20	3,0
18	105	20	3,1
19	110	20	3,1
20	115	20	3,2
21	120	20	3,2
22	125	20	3,2
23	130	20	3,3
24	100	25	2,8
25	105	25	2,8
26	110	25	2,9
27	115	25	2,9
28	120	25	3,0
29	125	25	3,0
30	130	25	3,1

Рис. 2.8. Расчетные значения y и соответствующие им значения x_1 и x_2 из примера 3

3. СТАТИСТИКА

В работе любого специалиста часто приходится сталкиваться с необходимостью обработки и анализа данных, полученных в результате наблюдения. Например, экономические данные формируются под действием множества факторов, не все из которых доступны внешнему контролю. Стохастическая природа экономических данных обусловливает необходимость применения специальных статистических методов для их анализа и обработки.

Или, например, изучаемые в ходе эксперимента некоторые психологические качества. Выборочное среднее значение как статистический показатель характеризует степень его развития в целом у той группы испытуемых, которая была подвергнута психодиагностическому обследованию. Сравнивая непосредственно средние значения двух или нескольких выборок, можно судить об относительной степени развития у людей, составляющих эти выборки, оцениваемого качества.

3.1. Основные понятия и определения

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.

Математическая статистика подразделяется на две основные области: **описательную** и **аналитическую** статистику. *Описательная* статистика охватывает методы описания статистических данных, представления их в форме таблиц, распределений и т. п.

Аналитическая статистика, или теория статистических выводов, ориентирована на обработку данных, полученных в ходе эксперимента, с целью формулировки выводов, имеющих прикладное значение для самых различных областей человеческой деятельности.

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, позволяющие проводить необходимый статистический анализ экономических, медико-биологических и иных типов данных.

При рассмотрении применения методов обработки статистических данных ограничимся только простейшими и наиболее часто используемыми методами, реализованными в **Мастере функций** и **Пакете анализа Excel**.

Выборочный метод

По охвату статистической совокупности исследование может быть сплошное или не сплошное. При сплошном статистическом исследовании группа наблюдения формируется путем полного охвата всех единиц изучаемого явления. Множество всех единиц наблюдения, охватываемых таким сплошным наблюдением, называется **генеральной совокупностью**.

Основным методом не сплошного наблюдения является выборочный метод. Если интересующая нас совокупность слишком многочисленна, либо ее элементы малодоступны, а также, если имеются другие причины (организационные, финансовые, физические и т. п.), не позволяющие изучать сразу все ее элементы, прибегают к изучению какой-то части этой совокупности. Эта выбранная для полного исследования группа элементов называется **выборкой** или **выборочной совокупностью**.

Выборка – это группа элементов, выбранная для исследования из всей совокупности элементов. Задача выборочного метода состоит в том, чтобы сделать правильные выводы относительно всего собрания объектов, их совокупности. Например, пробуя пищу, повар по одной ложке делает заключение о качестве приготавливаемого во всей кастрюле.

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.

Выборочная функция распределения

В практических задачах закон распределения случайных величин обычно неизвестен или известен с точностью до некоторых неизвестных параметров. В частности, невозможно рассчитать точное значение соответствующих вероятностей,

так как нельзя определить количество общих и благоприятных исходов. Поэтому вводится **статистическое определение вероятности.** По этому определению вероятность равна отношению числа испытаний (m), в которых событие появилось, к общему количеству произведенных испытаний (n). Такая вероятность называется **статистической частотой**.

В результате на практике сведения о законе распределения случайной величины получают независимыми многократными повторениями опыта, в котором измеряются значения интересующей исследователей случайной величины (варианты). На основе информации из полученной выборки можно построить приблизительные значения для функции распределения и других характеристик случайной величины.

Выборочной (эмпирической) функцией распределения случайной величины ξ, построенной по выборке $x_1, x_2, ..., x_n$, называется функция $F_n(x)$, равная доле таких значений x_i, что $x_i < x$, $i = 1, ..., n$.

Другими словами, $F_n(x)$ есть частота события $x_i < x$ в ряду $x_1, x_2, ..., x_n$.

Связь между эмпирической функцией распределения и функцией распределения (теоретической функцией распределения) такая же, как связь между частотой события и его вероятностью: функция $F_n(x) \to F(x)$ при $n \to \infty$.

Для построения выборочной функции распределения весь диапазон изменения случайной величины X разбивают на ряд интервалов одинаковой ширины. Число интервалов обычно выбирают не менее 5 и не более 15. Затем определяют число значений случайной величины X, попавших в каждый интервал. Поделив эти числа на общее количество наблюдений n, находят относительную частоту попадания случайной величины X в заданные интервалы. По найденным относительным частотам строят гистограммы выборочных функций распределения. Если соответствующие точки относительных частот соединить ломаной линией, то полученная диаграмма будет называться **полигоном частот.** Кумулятивная кривая будет получена, если по оси абсцисс откладывать интервалы, а по оси ординат – число или долю элементов совокупности, имеющих значение, меньшее или равное заданному.

При увеличении до бесконечности размера выборки выборочные функции распределения превращаются в теоретические: гистограмма превращается в график плотности распределения, а кумулятивная кривая – в график функции распределения.

В **Excel** для построения выборочных функций распределения 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. 3.1:

• 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.

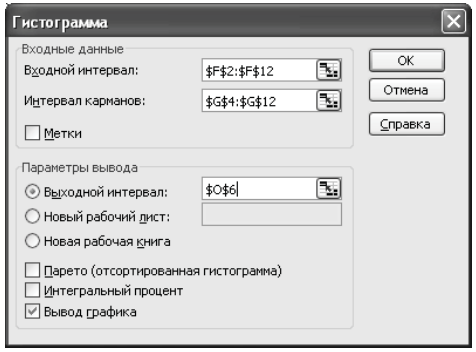


Рис. 3.1. Пример заполнения диалогового окна **Гистограмма**

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/$H$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 таблицу, представленную на рис. 3.2.

	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		

Рис. 3.2. Результат вычислений относительных и накопленных частот из примера 4

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 данных – 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 иметь такой вид, как на рис. 3.3.

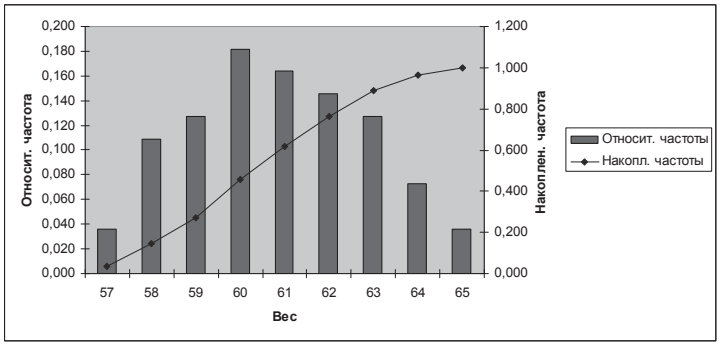


Рис. 3.3. Диаграмма относительных и накопленных частот из примера 4

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 (см. рис. 3.1):

- 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 **ОК**. В результате появляется таблица и диаграмма (рис. 3.4).

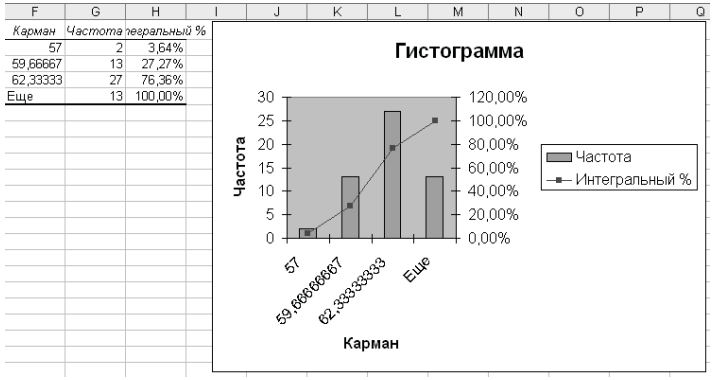


Рис. 3.4. Таблица и диаграмма из примера 5

Как видно, диаграмма на рис. 3.4 несколько отличается от диаграммы на рис. 3.3. Это объясняется тем, что диапазон карманов не был введен. Количество и границы интервалов определялись в процедуре **ГИСТОГРАММА** автоматически. Если бы в рабочее поле **Интервал карманов** был бы введен диапазон ячеек, определяющих выбранные интервалы, как в примере 4 (57, 58, 59, ..., 65), то полученная диаграмма была бы идентична предыдущей.

Выборочные характеристики

Замена теоретической функции распределения $F(x)$ на ее выборочный аналог $F_n(x)$ в определении математического ожидания, дисперсии, стандартного отклонения и т. п. приводят к выборочному среднему, выборочной дисперсии, выборочному стандартному отклонению и т. д. Выборочные характеристики являются оценками соответствующих характеристик генеральной совокупности. Эти оценки должны удовлетворять определенным требованиям. В соответствии с важнейшими требованиями оценки должны быть:

- несмещенными, то есть стремиться к истинному значению характеристики генеральной совокупности при неограниченном увеличении количества испытаний;
- состоятельными, то есть с ростом размера выборки оценка должна стремиться к значению соответствующего параметра генеральной совокупности с вероятностью, приближающейся к 1;
- эффективными, то есть для выборок равного объема используемая оценка должна иметь минимальную дисперсию.

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

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

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 $(n + 1)/2$. Если число членов ряда четное, то медиана равна среднему значению членов ряда с номерами $n/2$ и $n/2 + 1$.

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

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

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 по-

казатель 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 MX.

Выборочной квантилью называется решение уравнения

$$F_n(x) = p.$$

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.

3.2. Определение основных статистических характеристик средствами Мастера функций

В результате наблюдений или эксперимента получаются наборы данных, называемые **выборками**. Для проведения их анализа данные подвергаются статистической обработке. Первое, что всегда делается при обработке данных, это вычисление элементарных статистических характеристик выборок (как минимум: среднего, среднеквадратичного отклонения, ошибки среднего) по каждому параметру и по каждой группе. Полезно также вычислить эти характеристики для объединения родственных групп и суммарно по всем данным.

Использование специальных 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, ПЕРСЕНТИЛЬ.**

- 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.

- Функция **ПЕРСЕНТИЛЬ** позволяет получить квантили заданной выборки. Например, если ячейки A1:A7 содержат числа 10, 14, 5, 6, 10, 12 и 13, то квантилью со значением 0,1 является **ПЕРСЕНТИЛЬ(A1:A7;0,1),** равная 5,6.

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 ± 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		

3.3. Использование инструментов Пакета анализа для статистической обработки данных

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**.

Ввод данных. Исследуемые данные следует представить в виде таблицы, где столбцами являются соответствующие показатели. При создании таблицы **Excel** информация вводится в отдельные ячейки. Совокупность ячеек, содержащих анализируемые данные, называется входным диапазоном.

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.

Нахождение основных выборочных характеристик

Для определения характеристик выборки используется процедура **Описательная статистика.** Процедура позволяет получить статистический отчет, содержащий информацию о центральной тенденции и изменчивости входных данных. Для выполнения процедуры необходимо:

- выполнить команду **Сервис – Анализ данных**;
- в появившемся списке **Инструменты анализа** выбрать строку **Описательная статистика** и нажать кнопку **ОК** (рис. 3.5);

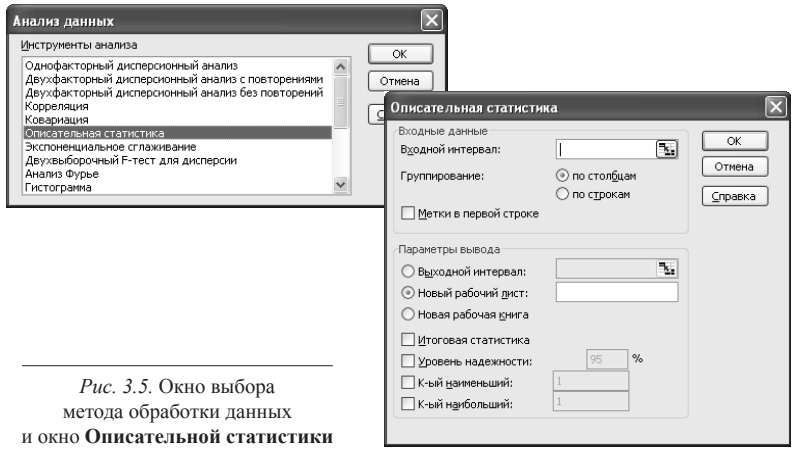


Рис. 3.5. Окно выбора метода обработки данных и окно **Описательной статистики**

• в появившемся диалоговом окне указать входной диапазон, то есть ввести ссылку на ячейки, содержащие анализируемые данные;

• указать выходной диапазон, то есть ввести ссылку на ячейки, в которые будут выведены результаты анализа;

• в разделе **Группировка** переключатель установить в положение **по столбцам**;

• установить флажок в поле **Итоговая статистика**;

• нажать кнопку **ОК**.

В результате анализа в указанном выходном диапазоне для каждого столбца данных выводятся следующие статистические характеристики: среднее, стандартная ошибка (среднего), медиана, мода, стандартное отклонение, дисперсия выборки, эксцесс, асимметричность, интервал, минимум, максимум, сумма, счет, наибольшее, наименьшее, уровень надежности.

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 таблица, представленная на рис. 3.6.

	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	

Рис. 3.6. Таблица из примера 7

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 **ОК**.

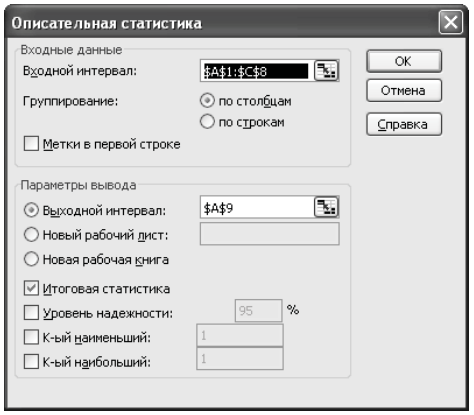


Рис. 3.7. Пример заполнения диалогового окна **Описательная статистика**

В результате анализа (рис. 3.8) 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	355,8089375	Стандартная ошибка	41,99277	Стандартная ошибка	243,5843
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

Рис. 3.8. Результаты работы инструмента **Описательная статистика**

Все полученные характеристики были рассмотрены ранее в разделе «Выборочные характеристики», за исключением последних четырех:

• **минимум** – значение минимального элемента выборки;

• **максимум** – значение максимального элемента выборки;
• **сумма** – сумма значений всех элементов выборки;
• **счет** – количество элементов в выборке.

Среди этих характеристик наиболее важными являются показатели Среднее, Стандартная ошибка (среднего) и Стандартное отклонение.

3.4. Принятие статистических решений

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 (рис. 3.9).

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.

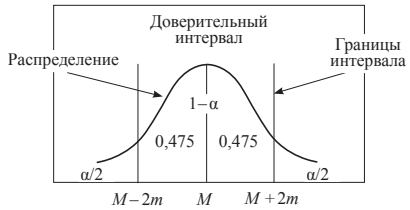


Рис. 3.9. 95-процентный доверительный интервал для среднего значения

Анализ одной выборки

Анализ однородности выборки. Одним из важных вопросов, возникающих при анализе выборки, является вопрос: относится та или иная варианта к данной статистической совокупности? Решение вопроса не представляет сложности, если распределение в этой совокупности является нормальным. Для этого достаточно использовать правило трех сигм. Согласно этому правилу в пределах $M \pm 3\sigma$ находится 99,7 % всех вариант. Поэтому если варианта попадает в этот интервал, то она считается принадлежащей к данной совокупности. Если не попадает, то она может быть отброшена. Хотя этот метод и предполагает нормальность исходного распределения, на практике он успешно работает и может быть использован в большинстве других случаев.

При числе элементов в выборке $n < 30$ способ более точного определения границ доверительного интервала по формуле

$$[M - t_{n,p}s;\ M + t_{n,p}s] \qquad (3.1)$$

будет показан ниже в примере 8. В формуле (3.1) M – среднее значение, s – стандартное отклонение, $t_{n,p}$ – табличное значение распределения Стьюдента с числом степеней свободы n и доверительной вероятностью p.

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 $\pm 2m$ (см. рис. 3.9). При доверительной вероятности 0,99 доверительный интервал составит примерно $\pm 3m$. Для более точного определения границ доверительного интервала можно воспользоваться формулой

$$\left[M - t_{n,p} \frac{s}{\sqrt{n}}; M + t_{n,p} \frac{s}{\sqrt{n}} \right],$$

где M – среднее значение;

s – стандартное отклонение;

$t_{n,p}$ – табличное значение распределения Стьюдента с числом степеней свободы n и доверительной вероятностью p;

n – количество элементов в выборке.

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*(1 – **альфа**) процентам, или, другими словами, **альфа**, равное 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 (рис. 3.10). Нажмите кнопку **ОК**.

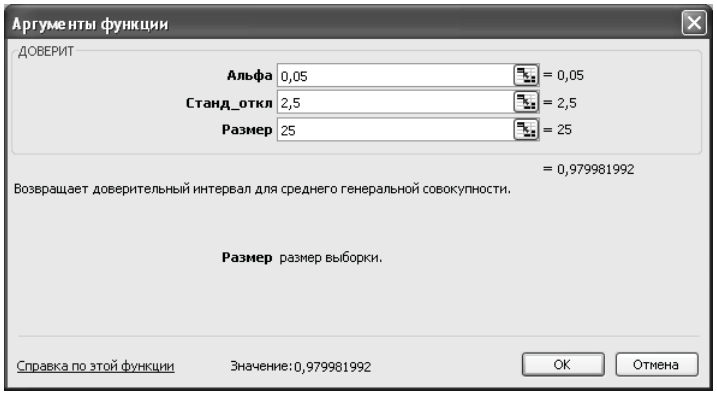


Рис. 3.10. Пример заполнения диалогового окна **ДОВЕРИТ**

4. В ячейке A1 появится полуширина 95-прцентного доверительного интервала для среднего значения выборки – 0,979981. Другими словами, с 95-процентным уровнем надежности можно утверждать, что средняя продолжительность разряда аккумулятора составляет 140 ± 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.

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 **ОК**.

4. В результате анализа в указанном выходном диапазоне для доверительной вероятности 0,95 получаем значения доверительного интервала (рис. 3.11).

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

Рис. 3.11. Исходная выборка (A1:A7) и результат вычислений (C3) для примера 9

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 ± 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 ± 9,975, считается принадлежащей данной совокупности с вероятностью 0,95. Выходящая за эти границы может быть отброшена с уровнем значимости $\alpha = 0{,}05$.

Проверка соответствия теоретическому распределению. 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 (рис. 3.12).

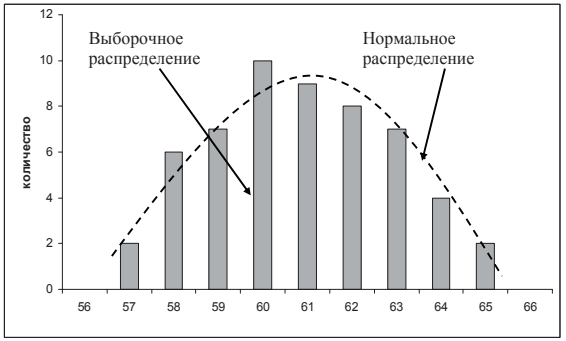


Рис. 3.12. Сопоставление выборочного распределения данных и кривой нормального распределения

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 H_0 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 использование графических и статистических (например, тест хи-квадрат или аналогичные) методов оценки, естественно дополняющих друг друга.

Использование критерия согласия хи-квадрат.

Для применения критерия желательно, чтобы объем выборки n был > 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. 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** – $J$13**; Интегральный** – 0. Получим в ячейке K4 0,033. Далее протягиванием скопируем содержимое ячейки K4 в диапазон ячеек K5:K12. Затем в ячейки L1 и L2 введем название нового столбца – **Теоретические частоты**. Установим курсор в ячейку L4 и введем формулу =$H$13*K4. Далее протягиванием скопируем содержимое ячейки L4 в диапазон ячеек L5:L12**.** Результаты вычислений представлены на рис. 3.13.

G	H	I	J	K	L
Вес	Абсолютные	Относительные	Накопленные	Теоретические	Теоретические
кг	частоты	частоты	частоты	частости	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		

Рис. 3.13. Результаты вычисления теоретических частостей и частот из примера 10

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 (рис. 3.14). В ячейке L13 появится значение вероятности того, что выборочные данные соответствуют нормальному закону распределения – 0,9842.

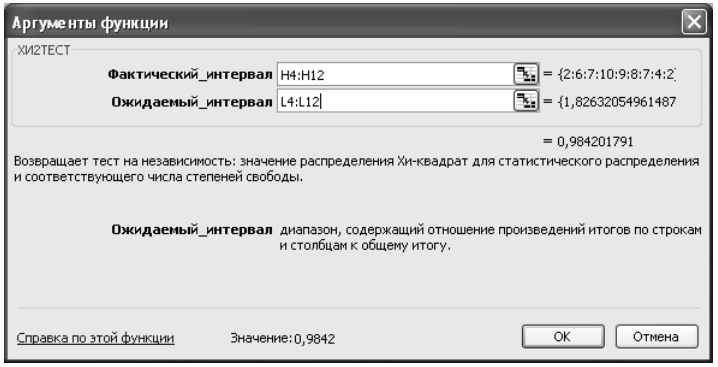


Рис. 3.14. Пример заполнения рабочих полей функции **ХИ2ТЕСТ**

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

дать, 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.

Анализ двух выборок

Выявление достоверности различий

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. 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, измеряется 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.

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

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, A11). 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.

В качестве упражнения рассмотрим пример.

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:

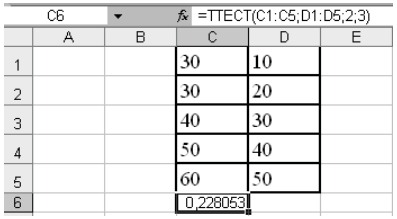
- группы состоят из различных студентов (тип 3);
- группы состоят из одних и тех же студентов, но первая – до посещения буфета, а вторая – после (тип 1).

Решение

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

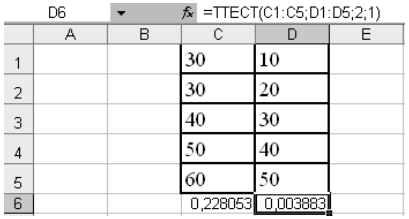
1. Установим курсор в свободную ячейку (например, C6). Вызовем **Мастер функций,** выберем категорию **Статистические** и функцию **ТТЕСТ**. В появившемся диалоговом окне функции **ТТЕСТ** введем исходные данные. Указателем

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 (без кавычек), 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.

Критерий 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:

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

где $F_{крит}$ 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.**

Непараметрические критерии. Непараметрические критерии используются в тех случаях, когда закон распределения данных отличается от нормального или неизвестен. Из большого числа непараметрических критериев рассмотрим критерий хи-квадрат.

Критерий согласия χ^2. 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). Критерий хи-квадрат относится к непараметрическим критериям.

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.

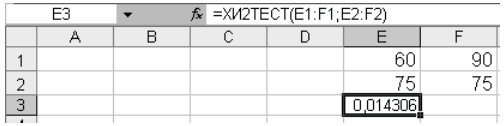
Пример 13. Пусть после окончания двух институтов экономического профиля трудоустроилось по специальности из первого института 90 человек, а из второго – 60 (обе группы молодых специалистов включали по 100 человек).

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.

Использование инструмента «Пакет анализа» для выявления различий между выборками

Для анализа двух выборок с помощью *t*-теста Стьюдента могут быть использованы следующие процедуры: *Парный двухвыборочный t-тест для средних; Двухвыборочный t-тест с одинаковыми дисперсиями и Двухвыборочный t-тест с различными дисперсиями.* Как указывалось в пункте «Анализ двух выборок», в общем случае необходимо воспользоваться процедурой *Двухвыборочный t-тест с различными дисперсиями*, так как процедуры *Парный двухвыборочный t-тест для средних* и *Двухвыборочный t-тест с одинаковыми дисперсиями* относятся к частным, специальным случаям.

Для выполнения процедуры анализа необходимо:

- выполнить команду **Сервис – Анализ данных**;
- в появившемся списке **Инструменты анализа** выбрать строку **Двухвыборочный *t*-тест с различными дисперсиями**, щелкнуть по кнопке **ОК**;
- в появившемся диалоговом окне указать **Интервал переменной 1**, то есть ввести ссылку на первый диапазон анализируемых данных, содержащий один столбец данных;
- указать **Интервал переменной 2**, то есть ввести ссылку на второй диапазон анализируемых данных, содержащий один столбец данных;
- указать **Выходной диапазон**;
- нажать кнопку **ОК**.

Результаты анализа. В выходной диапазон будут выведены: средняя, дисперсия и число наблюдений для каждой переменной, гипотетическая разность средних, *df* (число степеней свободы), значение *t*-статистики, $P(T <= t)$ одностороннее, *t* критическое одностороннее, $P(T <= t)$ двухстороннее, t критическое двухстороннее.

Интерпретация результатов. Если величина вероятности случайного появления анализируемых выборок ($P(T <= t)$ двухстороннее) меньше уровня значимости ($\alpha = 0{,}05$), принято считать, что различия между выборками не случайные, то есть различия достоверные.

Пример 14. Рассматривается заработная плата обслуживающего персонала и работников ресторана гостиницы (из примера 7).

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 **ОК**.

Результаты анализа. В выходном диапазоне C1:E13 появятся результаты процедуры **Двухвыборочный *t*-тест с различными дисперсиями** (рис. 3.15).

	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	

Рис. 3.15. Исходные данные (A1:B8) и результаты анализа (C1:E13)

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 ($P(T <= t)$ 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.

3.5. 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.

Влияние случайной составляющей характеризует внутригрупповая дисперсия, а влияние изучаемого фактора – межгрупповая. Внутригрупповая дисперсия рассчитывается по формуле:

$$s_2^2 = \frac{1}{m(n-1)} \sum_{i=1}^{n} \sum_{j=1}^{m} (x_{ij} - M_i)^2,$$

межгрупповая:

$$s_1^2 = \frac{1}{m-1} \sum_{i=1}^{m} (M_i - M)^2,$$

$$M_i = \frac{1}{n} \sum_{j=1}^{n} x_{ij} \ .$$

Здесь M – общее среднее, m – количество групп, n – количество элементов в группе.

В **MS Excel** для проведения однофакторного дисперсионного анализа используется процедура **Однофакторный дисперсионный анализ**.

Для проведения дисперсионного анализа необходимо:

• ввести данные в таблицу так, чтобы в каждом столбце оказались данные, соответствующие одному значению исследуемого фактора, а столбцы располагались в порядке возрастания (убывания) величины исследуемого фактора;

• выполнить команду **Сервис – Анализ данных**;

• в появившемся диалоговом окне **Анализ данных** в списке **Инструменты анализа** выбрать процедуру **Однофакторный дисперсионный анализ**;

• в появившемся диалоговом окне задать **Входной интервал**, то есть ввести ссылку на диапазон анализируемых данных, содержащий все столбцы данных (рис. 3.16);

• в разделе **Группировка** переключатель установить в положение **по столбцам**;

• указать выходной диапазон, то есть ввести ссылку на ячейки, в которые будут выведены результаты анализа. Для этого следует поставить переключатель в положение **Выходной интервал**, навести указатель мыши на левую верхнюю ячейку выходного диапазона и щелкнуть левой кнопкой мыши. Размер выходного диапазона будет определен автоматически, и на экран будет выведено сообщение в случае возможного наложения выходного диапазона на исходные данные;

• нажать кнопку **ОК**.

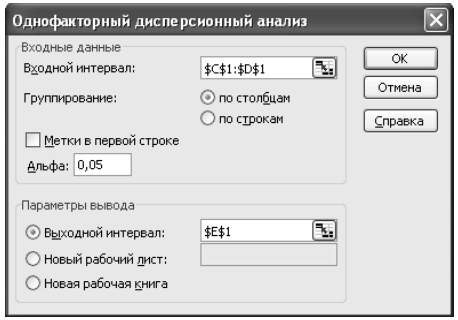


Рис. 3.16. Пример заполнения диалогового окна **Однофакторный дисперсионный анализ**

Результаты анализа. Выходной диапазон будет включать в себя результаты дисперсионного анализа: средние, дисперсии, критерий Фишера и другие показатели.

Интерпретация результатов. Влияние исследуемого фактора определяется по величине значимости критерия Фишера, которая находится в таблице **Дисперсионный анализ** на пересечении строки **Между группами** и столбца **Р-Значение**. В случаях, когда **Р-Значение** < 0,05, критерий Фишера значим и влияние исследуемого фактора можно считать доказанным.

Кроме рассмотренной процедуры однофакторного дисперсионного анализа для проведения двухфакторного дисперсионного анализа в пакете анализа реализованы процедуры **Двухфакторный дисперсионный анализ с повторениями** и **Двухфакторный дисперсионный анализ без повторений**.

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	87	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 ОК.

Результаты анализа. В результате будет получена таблица, показанная на рис. 3.17.

	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				

Рис. 3.17. Результат работы инструмента **Однофакторный дисперсионный анализ**

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.

3.6. Корреляционный анализ

Важным разделом статистического анализа является корреляционный анализ, служащий для выявления взаимосвязей между выборками.

Коэффициент корреляции

Выявление взаимосвязей. 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 **корреляцию** и **регрессию**.

Регрессионный анализ (см. раздел «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, r) – параметр, характеризующий степень линейной взаимосвязи между двумя выборками. Коэффициент корреляции изменяется от -1 (строгая обратная линейная зависимость) до 1 (строгая прямая пропорциональная 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 постоянном объеме). 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.

Выборочный коэффициент линейной корреляции между двумя случайными величинами X и Y рассчитывается по формуле

$$r = \frac{\sum (x - M_x)(y - M_y)}{\sqrt{\sum (x - M_x)^2 (y - M_y)^2}}.$$

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

На практике коэффициент корреляции принимает некоторые промежуточные значения между 1 и -1 (рис. 3.18). 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 коэффициент 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.

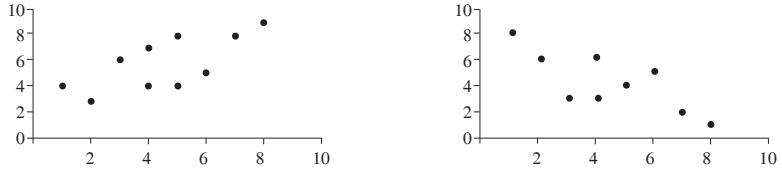


Рис. 3.18. Примеры прямой (r = 0,7, а) и обратной (r = -0,8, б) корреляционной зависимости

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

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

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**. 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).

Корреляционная матрица

При большом числе наблюдений, когда коэффициенты корреляции необходимо последовательно вычислять из нескольких рядов числовых данных, для удобства получаемые коэффициенты сводят в таблицы, называемые корреляционными матрицами.

Корреляционная матрица – это квадратная (или прямоугольная) таблица, в которой на пересечении соответствующих строки и столбца находится коэффициент корреляции между соответствующими параметрами.

В **MS Excel** для вычисления корреляционных матриц используется процедура **Корреляция**. Процедура позволяет получить корреляционную матрицу, содержащую коэффициенты корреляции между различными параметрами.

Для реализации процедуры необходимо:

- выполнить команду **Сервис – Анализ данных**;
- в появившемся списке **Инструменты анализа** выбрать строку **Корреляция** и нажать кнопку **ОК**;
- в появившемся диалоговом окне указать **Входной интервал**, то есть ввести ссылку на ячейки, содержащие анализируемые данные. Входной интервал должен содержать не менее двух столбцов;
- в разделе **Группировка** переключатель установить в соответствии с введенными данными (например, **по столбцам**);
- указать выходной диапазон, то есть ввести ссылку на ячейки, в которые будут выведены результаты анализа. Для этого следует установить флажок **Выходной интервал**, далее навести указатель мыши на правое поле ввода **Выходной интервал** и щелкнуть левой кнопкой мыши, затем указатель мыши навести на левую верхнюю ячейку выходного диапазона и щелкнуть левой кнопкой мыши. Размер выходного диапазона будет определен автоматически, и на экран будет выведено сообщение в случае возможного наложения выходного диапазона на исходные данные (рис. 3.19);
- нажать кнопку **ОК**.

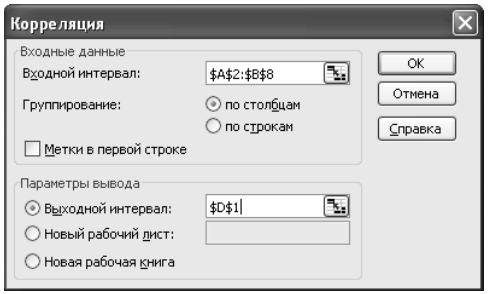


Рис. 3.19. Пример установки параметров корреляционного анализа

Результаты анализа. В выходной диапазон будет выведена корреляционная матрица, в которой на пересечении каждых строки и столбца находится коэффициент корреляции между соответствующими параметрами. Ячейки выходного диапазона, имеющие совпадающие координаты строк и столбцов, содержат значение 1, так как каждый столбец во входном диапазоне полностью коррелирует с самим собой.

Интерпретация результатов. Рассматривается отдельно каждый коэффициент корреляции между соответствующими параметрами. Его числовое значение оценивается по эмпирическим правилам, изложенным в пункте «Коэффициент корреляции». Отметим, что, хотя в результате будет получена треугольная матрица, корреляционная матрица симметрична, и коэффициенты корреляции $r_{ij} = r_{ji}$.

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

Для выполнения корреляционного анализа введите в диапазон A1:G3 исходные данные (рис. 3.20).

	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**

Рис. 3.20. Исходные данные из примера 17

Затем выполните команду **Сервис – Анализ** данных и выберите строку **Корреляция**. В появившемся диалоговом окне укажите **Входной интервал** B1: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 A4. Нажмите кнопку **ОК**.

Результаты анализа. В выходном диапазоне получаем корреляционную матрицу (рис. 3.21).

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

Рис. 3.21. Результаты вычисления корреляционной матрицы из примера 17

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).

3.7. Регрессионный анализ

При исследовании взаимосвязей между выборками помимо корреляции различают также и **регрессию**. Регрессия используется для анализа воздействия на отдельную зависимую переменную значений одной или более независимых переменных. Соответственно, наряду с корреляционным анализом еще одним инструментом изучения стохастических зависимостей является регрессионный анализ. Регрессионный анализ устанавливает формы зависимости между случайной величиной Y (зависимой) и значениями одной или нескольких переменных величин (независимых), причем значения последних считаются точно заданными. Такая зависимость обычно определяется некоторой математической моделью (уравнением регрессии), содержащей несколько неизвестных параметров. В ходе регрессионного анализа на основании выборочных данных находятся оценки этих параметров, определяются статистические ошибки оценок или границы доверительных интервалов и проверяется соответствие (адекватность) принятой математической модели экспериментальным данным.

В линейном регрессионном анализе связь между случайными величинами предполагается линейной. В самом простом случае в линейной регрессионной модели имеются две переменные X и Y. И требуется по n парам наблюдений (X_1, Y_1), (X_2, Y_2), ..., (X_n, Y_n) построить (подобрать) прямую линию, называемую линией регрессии, которая наилучшим образом приближает наблюдаемые значения. Урав-

нение этой линии $Y = aX + b$ является регрессионным уравнением. С помощью регрессионного уравнения можно предсказать ожидаемое значение зависимой величины Y_0, соответствующее заданному значению независимой переменной X_0.

Таким образом, можно сказать, что линейный регрессионный анализ заключается в подборе графика и его уравнения для набора наблюдений. В регрессионном анализе все признаки (переменные), входящие в уравнение, должны иметь непрерывную, а не дискретную природу.

В случае, когда рассматривается зависимость между одной зависимой переменной Y и несколькими независимыми переменными $X_1, X_2, ..., X_n$, говорят о множественной линейной регрессии. В этом случае регрессионное уравнение имеет вид

$$Y = a_0 + a_1X_1 + a_2X_2 + ... + a_nX_n,$$

где $a_1, a_2, ..., a_n$ – требующие определения коэффициенты при независимых переменных $X_1, X_2, ... X_n$;

a_0 – константа.

Мерой эффективности регрессионной модели является коэффициент детерминации R^2 (R-квадрат). Коэффициент детерминации (R-квадрат) определяет, с какой степенью точности полученное регрессионное уравнение описывает (аппроксимирует) исходные данные.

Исследуется также значимость регрессионной модели с помощью F-критерия (Фишера). Если величина F-критерия значима ($p < 0,05$), то регрессионная модель является значимой.

Достоверность отличия коэффициентов $a_0, a_1, a_2, ..., a_n$ от нуля проверяется с помощью критерия Стьюдента. В случаях, когда $p > 0,05$, коэффициент может считаться нулевым, а это означает, что влияние соответствующей независимой переменной на зависимую переменную недостоверно, и эта независимая переменная может быть исключена из уравнения.

В **MS Excel** экспериментальные данные аппроксимируются линейным уравнением до 16 порядка:

$$Y = a_0 + a_1X_1 + a_2X_2 + ... + a_{16}X_{16},$$

где Y – зависимая переменная,

$X_1, ..., X_{16}$ – независимые переменные,

$a_0, a_1..., a_{16}$ – искомые коэффициенты регрессии.

Для получения коэффициентов регрессии используется процедура **Регрессия** из **Пакета анализа**. Кроме того, могут быть использованы функция **ЛИНЕЙН** для получения параметров регрессионного уравнения и функция **ТЕНДЕНЦИЯ** для получения предсказанных значений Y в требуемых точках.

Для реализации процедуры **Регрессия** необходимо:

- выполнить команду **Сервис – Анализ данных**;
- в появившемся диалоговом окне **Анализ данных** в списке **Инструменты анализа** выбрать строку **Регрессия**;
- в появившемся диалоговом окне задать **Входной интервал Y,** то есть ввести ссылку на диапазон анализируемых зависимых данных, содержащий один столбец данных;

- указать **Входной интервал X,** то есть ввести ссылку на диапазон независимых данных, содержащий до 16 столбцов анализируемых данных;
- указать выходной диапазон, то есть ввести ссылку на ячейки, в которые будут выведены результаты анализа (рис. 3.22);

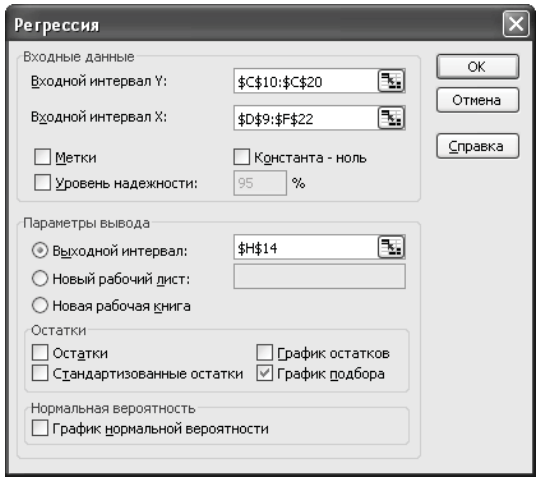


Рис. 3.22. Пример заполнения диалогового окна **Регрессия**

- если необходимо визуально проверить отличие экспериментальных точек от предсказанных по регрессионной модели, следует установить флажок в поле **График подбора**;
- нажать кнопку **ОК**.

Результаты анализа. Выходной диапазон будет включать в себя результаты дисперсионного анализа, коэффициенты регрессии, стандартную погрешность вычисления Y, среднеквадратичные отклонения, число наблюдений, стандартные погрешности для коэффициентов.

Интерпретация результатов. Значения коэффициентов регрессии находятся в столбце **Коэффициенты** и соответствуют:

- Y-пересечение – a_0;
- переменная X1 – a_1
- переменная X2 – a_2 и т. д.

В столбце **P-Значение** приводится достоверность отличия соответствующих коэффициентов от нуля. В случаях, когда $P > 0,05$, коэффициент может считаться нулевым; это означает, что соответствующая независимая переменная практически не влияет на зависимую переменную.

Приводимое значение **R-квадрат** (коэффициент детерминации) определяет, с какой степенью точности полученное регрессионное уравнение аппроксимирует

исходные данные. Если R-квадрат > 0,95, говорят о высокой точности аппроксимации (модель хорошо описывает явление). Если R-квадрат лежит в диапазоне от 0,8 до 0,95, говорят об удовлетворительной аппроксимации (модель в целом адекватна описываемому явлению). Если R-квадрат < 0,6, принято считать, что точность аппроксимации недостаточна и модель требует улучшения (введения новых независимых переменных, учета нелинейностей и т. д.).

Пример 18. В отделе снабжения гостиницы имеется информация об изменении стоимости стирального порошка за длительный период времени. Сопоставляя ее с изменениями курса доллара за этот же период времени, можно построить регрессионное уравнение. Ниже приведены стоимость пачки стирального порошка (в рублях) и соответствующий курс доллара (руб./USD).

Номер	1	2	3	4	5	6	7	8
Порошок	5	7	9	12	15	16	20	25
Курс	6,3	9	12	15	19	21	25	29,3

Необходимо на основании этих данных построить регрессионное уравнение, позволяющее по курсу доллара определять предполагаемую стоимость пачки стирального порошка.

Решение

1. Введите данные в рабочую таблицу: стоимость пачки порошка – в диапазон A1:A8; курс доллара – в диапазон B1:B8.

2. Выполните команду **Сервис – Анализ данных** и выберите строку **Регрессия**.

3. В появившемся диалоговом окне (рис. 3.22) задайте **Входной интервал Y** – это диапазон ячеек A1:A8 (обратите внимание, что зависимые данные – это те данные, которые предполагается вычислять).

4. Также укажите **Входной интервал X,** задав диапазон независимых данных B1:B8 (независимые данные – это те данные, которые будут измеряться или наблюдаться).

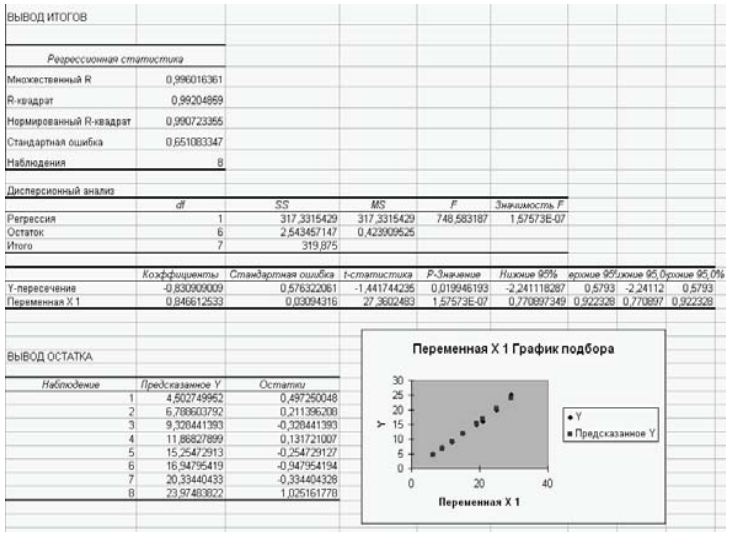
5. Установите флажок в поле **График подбора**.

6. Далее укажите **Выходной диапазон,** например, ячейку C1.

7. Нажмите кнопку **ОК**.

Результаты анализа. В выходном диапазоне появятся следующие результаты и график подбора (рис. 3.23).

Интерпретация результатов. В таблице **Дисперсионный анализ** оценивается общее качество полученной модели: ее достоверность по уровню значимости критерия Фишера – p, который должен быть меньше, чем 0,05 (строка **Регрессия**, столбец **Значимость F**, в примере – 1,58E-07 (0,000000158), то есть p = 0,000000158 и модель значима) и степень точности описания моделью процесса – **R-квадрат** (вторая строка сверху в таблице **Регрессионная статистика**, в примере R-квадрат = 0,992). Поскольку R-квадрат > 0,95, можно говорить о высокой точности аппроксимации (модель хорошо описывает явление (рис. 3.23)).



ВЫВОД ИТОГОВ

Регрессионная статистика

Множественный R	0,996016361
R-квадрат	0,99204859
Нормированный R-квадрат	0,990723355
Стандартная ошибка	0,651083347
Наблюдения	8

Дисперсионный анализ

	df	*SS*	*MS*	*F*	*Значимость F*
Регрессия	1	317,3315429	317,3315429	748,583187	1,57573E-07
Остаток	6	2,543457147	0,423909525		
Итого	7	319,875			

	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	-0,830909009	0,576322061	-1,441744235	0,019946193	-2,241118287	0,5793	-2,24112	0,5793
Переменная X 1	0,846612533	0,03094316	27,3602483	1,57573E-07	0,770997349	0,922328	0,770997	0,922328

ВЫВОД ОСТАТКА

Наблюдение	*Предсказанное Y*	*Остатки*
1	4,502749962	0,497250048
2	6,788603792	0,211396208
3	9,328441393	-0,328441393
4	11,86827899	0,131721007
5	15,25472913	-0,254729127
6	16,94795419	-0,947954194
7	20,33440433	-0,334404328
8	23,97483822	1,025161778

Рис. 3.23. Результаты анализа и график соответствия экспериментальных точек и предсказанных по регрессионной модели из примера 18

Далее необходимо определить значения коэффициентов модели. Они определяются из таблицы в столбце **Коэффициенты** – в строке **Y-пересечение** приводится свободный член; в строках соответствующих переменных приводятся значения коэффициентов при этих переменных. В столбце **p-значение** приводится достоверность отличия соответствующих коэффициентов от нуля. В случаях, когда p > 0,05, коэффициент может считаться нулевым. Это означает, что соответствующая независимая переменная практически не влияет на зависимую переменную и коэффициент может быть убран из уравнения.

Отсюда выражение для определения стоимости пачки порошка в рублях будет иметь следующий вид: **-0,83 + 0,847*(Курс доллара, руб./USD).**

Полученная модель с высокой точностью позволяет определять стоимость пачки стирального порошка (R^2 = 99,2 %).

Воспользовавшись полученным уравнением, можно рассчитать ожидаемую стоимость пачки стирального порошка при изменениях курса доллара. Например, при курсе доллара 35 руб./USD ожидаемая стоимость пачки порошка равна 28,8 руб.

Пример 19. Построить регрессионную модель для предсказания изменений уровня заболеваемости органов дыхания (*Y)* в зависимости от содержания в воздухе двуокиси углерода (X_1) и степени запыленности (X_2). В таблице приведены данные наблюдений в течение 29 месяцев.

X_1	X_2	Y
1	1,3	1160
1	1,3	1155
1,1	1,4	1158
1,1	1,4	1157
1,1	1,5	1160
1,1	1,5	1161
1	1,4	1157
1	1,5	1159
1,2	1,6	1256
1,2	1,7	1260
0,6	1	1040
0,6	1	1039
0,7	1,1	1039
0,7	1,15	1040
0,75	1,2	1040
0,7	1,2	1039
0,7	1,3	1040
0,7	1,3	1039
0,8	1,4	1140
0,8	1,4	1138
0,78	1,5	1240
0,8	1,5	1239
0,78	1,5	1241
0,78	1,6	1240
0,8	1,7	1239
0,8	1,8	1239
0,75	1,8	1240
0,78	1,9	1238
0,75	1,9	1238

Решение

1. Введите данные наблюдений в диапазон A1:C30 рабочей таблицы **Excel**.

2. Выполните команду **Сервис – Анализ данных** и выберите строку **Регрессия**.

3. В появившемся диалоговом окне зададим **Входной интервал Y** – это диапазон ячеек C2:C30.

4. Также укажем **Входной интервал X** – это диапазон независимых данных A2:B30.

5. Установите флажок в поле **График подбора**.

6. Далее укажите **Выходной диапазон** (например, ячейку D1). Нажмите кнопку **ОК**.

7. В выходном диапазоне появятся результаты регрессионного анализа и графики предсказанных точек (рис. 3.24).

D	E	F	G	H	I	J	K	L
ВЫВОД ИТОГОВ								
Регрессионная статистика								
Множественный R	0,889678534							
R-квадрат	0,791527894							
Нормированный R-квадрат	0,775491579							
Стандартная ошибка	39,29298772							
Наблюдения	29							
Дисперсионный анализ								
	df	*SS*	*MS*	*F*	*ачимость F*			
Регрессия	2	152412,9	76206,45	49,35846	1,4E-09			
Остаток	26	40142,41	1543,939					
Итого	28	192555,3						
	Коэффициенты	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
Y-пересечение	681,907804	50,29342	13,55859	2,67E-13	578,5282	785,2874	578,5282	785,2874
Переменная X 1	90,90811	43,89337	2,071112	0,04841	0,683991	181,1322	0,683991	181,1322
Переменная X 2	274,6664036	31,83362	8,628186	4,16E-09	209,2315	340,1014	209,2315	340,1014

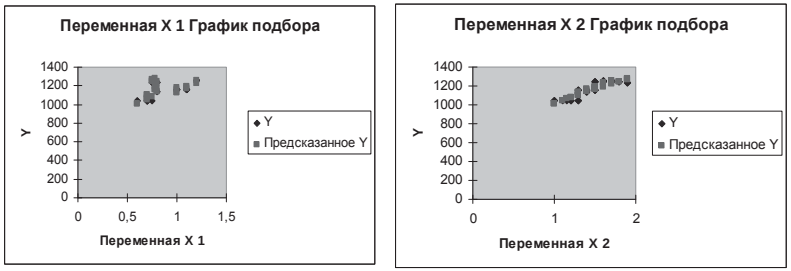


Рис. 3.24. Графики расположения фактических и предсказанных точек (пример 19)

Интерпретация результатов. В таблице **Дисперсионный анализ** оценивается достоверность полученной модели по уровню значимости критерия Фишера (строка **Регрессия**, столбец **Значимость F**, в примере – 1,4E-09 ($1{,}4*10^{-9}$), то есть $p < 0{,}05$ и модель значима) и степень описания моделью процесса – R-квадрат (вторая строка сверху в таблице **Регрессионная статистика**, в примере R-квадрат = 0,79). Можно говорить о довольно высокой точности аппроксимации (модель хорошо описывает зависимость заболеваемости от содержания углекислого газа и запыленности воздуха (рис. 3.24)).

Далее необходимо определить значения коэффициентов модели. Они определяются из таблицы в столбце **Коэффициенты** – в строке **Y-пересечение** приво-

дится свободный член a_0 = 682; в строках соответствующих переменных приводятся значения коэффициентов при этих переменных a_1 = 91 и a_2 = 275. В столбце **p-значение** приводится достоверность отличия соответствующих коэффициентов от нуля. Все коэффициенты значимы, то есть $p < 0{,}05$, и коэффициенты могут считаться не равными нулю.

Поэтому выражение для определения уровня заболеваемости органов дыхания в зависимости от содержания углекислого газа и пыли в воздухе будет иметь вид:

$$\mathbf{Y = 682 + 91X_1 + 275*X_2.}$$

3.8. Уравнение регрессии

3.8.1. Пример построения уравнения регрессии (линейная модель)

Рассмотрим более подробно средства **MS Excel** для построения уравнения регрессии.

Пусть Вы являетесь менеджером фирмы по продажам подержанных автомобилей и постоянно ведете учет проданных автомобилей. В вашем распоряжении имеются две наблюдаемые величины: x – номер недели, y – число проданных за неделю автомобилей (табл. 3.1). Фирма совсем молодая, была создана шесть недель назад, и поэтому в вашем распоряжении имеется статистика только за этот весьма ограниченный промежуток времени.

Таблица 3.1

Значения наблюдаемых величин

Наблюдаемые величины	Значения					
x	1	2	3	4	5	6
y	7	9	12	13	14	17

Вы хотите сначала смоделировать ту динамику продаж, которая имеет место, а на основе построенной модели затем попытаться заглянуть в будущее, т. е. спрогнозировать ожидаемый объем продаж на ближайшие недели.

В качестве модели вы решили взять простейшую модель $\boldsymbol{y = mx + b}$, наилучшим образом описывающую наблюдаемые значения. Обычно m и b подбираются так, чтобы минимизировать сумму квадратов разностей теоретических и наблюдаемых значений зависимой переменной (y), т. е. минимизировать:

$$z = \sum_{i=1}^{n} \left(y_i - mx_i - b\right)^2,$$

где n – число наблюдений (в данном случае n = 6).

Для решения этой задачи необходимо выполнить следующие действия:

1. Заполнить ячейки A2:B7 (рис. 3.25).

2. Отвести под переменные m и b ячейки D2 и E2.

3. В ячейку F2 ввести минимизируемую функцию (это формула массива, поэтому не забудьте завершить ее ввод нажатием комбинации клавиш <Shift> + <Ctrl> + <Enter>).

{=СУММ ((B2: B7 – D2 * A2 : A7 – E2) ^ 2)}

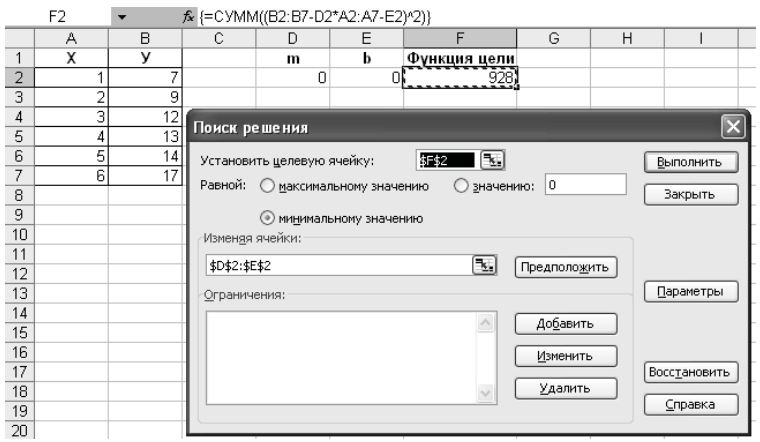
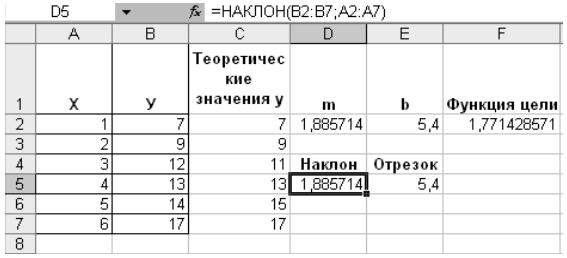


Рис. 3.25. Исходные данные для построения линейной модели и диалоговое окно **Поиск решения**

4. Выполнить команду **Сервис – Поиск решения**. Диалоговое окно **Поиск решения** следует заполнить, как показано на рис. 3.25. Отметим, что на переменные m и b не налагается никаких ограничений.

5. Нажать кнопку **Выполнить**. В результате вычислений средство **Поиск решения** найдет m = 1,88571 и b = 5,40 (см. рис. 3.26).



	A	B	C	D	E	F
1	X	У	Теоретические значения y	m	b	Функция цели
2	1	7	7	1,885714	5,4	1,771428571
3	2	9	9			
4	3	12	11	Наклон	Отрезок	
5	4	13	13	1,885714	5,4	
6	5	14	15			
7	6	17	17			
8						

Рис. 3.26. Теоретическое значение наблюдаемой величины и коэффициенты уравнения регрессии

3.8.2. Функции MS Excel для построения линейного уравнения регрессии

Рассмотрим некоторые функции категории **Статистические** для построения уравнения регрессии.

Параметры m и b линейной модели $\boldsymbol{y = mx + b}$ из предыдущего примера можно определить при помощи функций **НАКЛОН** и **ОТРЕЗОК**.

Функция **НАКЛОН** определяет коэффициент наклона линейного тренда, а функция **ОТРЕЗОК** – точку пересечения линии линейного тренда с осью ординат.

Синтаксис:

НАКЛОН (изв_знач_у; изв_знач_х)

ОТРЕЗОК (изв_знач_у; изв_знач_х)

Здесь:

• **изв_знач_у** – массив известных значений зависимой наблюдаемой величины;

• **изв_знач_х** – массив известных значений независимой наблюдаемой величины. Если опущены **изв_знач_х**, то предполагается, что это массив {1; 2; 3;…} такого же размера, как и **изв_знач_у**.

Функции **НАКЛОН** и **ОТРЕЗОК** вычисляют результат по следующим формулам:

$$m = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2},$$

$$b = y_{av} - m x_{av},$$

где $x_{av} = \frac{\sum_{i=1}^{n} x_i}{n}, y_{av} = \frac{\sum_{i=1}^{n} y_i}{n}$.

В ячейках D5 и E5 (рис. 3.26) найдены значения *m* и *b*, соответственно по формулам:

=НАКЛОН (B2:B7; A2:A7)

=ОТРЕЗОК (B2:B7; A2:A7)

Найдя коэффициенты уравнения регрессии, на их основе легко определить теоретические значения наблюдаемой величины. Для этого:

1. Введите в ячейку C2 формулу:

= $D$5*A2+$E$5

2. Выберите ячейку C2, расположите указатель мыши на маркере автозаполнения и протяните его вдоль диапазона C2:C7.

Теоретическое значение можно также вычислить с помощью функции **ПРЕДСКАЗ**, не определяя предварительно коэффициенты линейной модели, в фиксированной точке.

Синтаксис:

ПРЕДСКАЗ (х; изв_знач_у; изв_знач_х).

Здесь:

• **х** – точка данных, для которой предсказывается значение;

• **изв_знач_у –** массив известных значений зависимой наблюдаемой величины;

• **изв_знач_х** – массив известных значений независимой наблюдаемой величины. Если **изв_знач_х** опущены, то предполагается, что это массив {1; 2; 3; …} такого же размера, как и **изв_знач_у**.

Например, теоретическое значение в ячейке C2 можно было бы также определить по формуле:

= ПРЕДСКАЗ (A2; $B$2:$B2$7; $A$2:$A$7)

Функция **ТЕНДЕНЦИЯ** (подробно см. раздел 2) вычисляет значения уравнения линейной регрессии для целого диапазона значений независимой переменной как для случайного одномерного, так и для многомерного уравнения регрессии. Многомерная линейная модель регрессии имеет вид:

$$\boldsymbol{y = m_1 x_1 + \ldots + m_n x_n + b}.$$

Функция **ЛИНЕЙН** возвращает массив $\{m_n, ..., m_1, b\}$ значений параметров уравнения многомерной линейной регрессии.

3.8.3. Пример построения линейного уравнения регрессии и линии тренда

В двух предыдущих пунктах было показано, как находить коэффициенты уравнения регрессии. Теперь можно построить его диаграмму. В **MS Excel** линия уравнения регрессии называется линией тренда, которая показывает тенденцию изменения данных и служит для составления прогнозов. Для создания линии тренда на основе диаграммы используется один из пяти типов аппроксимаций или линейная фильтрация (табл. 3.2).

Таблица 3.2

Типы аппроксимаций

Тип	Описание
Линейная	$y = mx + b$ где *m* – тангенс угла наклона, *b* – точка пересечения с осью ординат
Логарифмическая	$y = m\ln x + b$ где *m* и *b* – константы
Полиномиальная	$y = m_6 x^6 + \ldots + m_1 x + b$ где $m_6 \ldots, m_1$ и *b* – константы
Степенная	$y = bx^m$ где *m* и *b* – константы
Экспоненциальная	$y = bm^x$ где *m* и *b* – константы
Линейная фильтрация	Каждая точка данных на линии тренда строится на основе среднего указанного числа точек данных (периодов). Чем больше число периодов устанавливается, тем более гладкой, но менее точной становится линия тренда

На диаграмме можно выделить любой ряд данных и добавить к нему линию тренда. Когда линия тренда добавляется к ряду данных, она связывается с ним, и поэтому при изменении значений любых точек ряда данных линия тренда автоматически пересчитывается и обновляется на диаграмме. Кроме того, имеется возможность выбирать точку, в которой линия тренда пересекает ось ординат, добавлять к диаграмме уравнение регрессии и величину достоверности аппроксимации.

Покажем на нашем примере по продажам автомобилей (см. табл. 3.1 и рис. 3.25), как строится линия тренда. Для этого:

1. При помощи **Мастера диаграмм** постройте по диапазону ячеек A2:B7 точечную диаграмму.

2. Из контекстного меню выберите команду **Добавить линию тренда.** На экране отобразится диалоговое окно **Линия тренда.**

3. На вкладке **Тип** диалогового окна **Линия тренда** выберете тип линии тренда. В данном случае – **Линейная** (рис. 3.27).

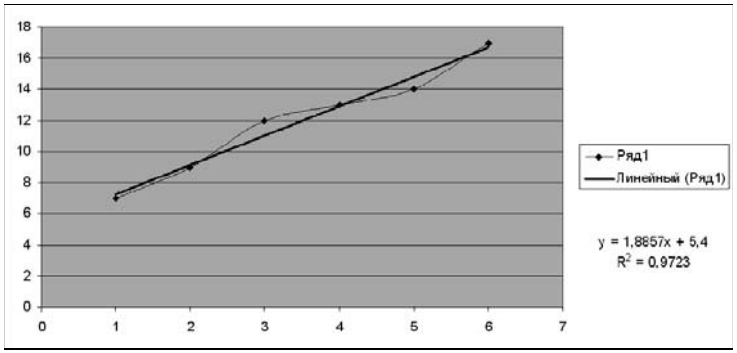


Рис. 3.27. График и **Линия тренда** для примера по продажам автомобилей

4. Заметим, что при выводе линии тренда можно показать величину достоверности аппроксимации, т. е. квадрат коэффициента корреляции (R^2). По коэффициенту корреляции можно судить о правомерности использования линейного уравнения регрессии. Если он лежит в диапазоне от 0,9 до 1, то данную зависимость можно использовать для предсказания результата. Чем коэффициент корреляции ближе к единице, тем он более обоснованно указывает на линейную зависимость между наблюдаемыми величинами. Если коэффициент корреляции лежит близко к -1, то это говорит об обратной зависимости между ними.

Результат выполнения команды **Добавить линию тренда** приведен на рис. 3.27. Квадрат коэффициента корреляции равен 0,9723. Следовательно, линейная модель может быть использована для предсказания результатов.

3.8.4. Экспоненциальная модель

Другой часто встречающейся на практике регрессионной моделью является экспоненциальная, которая описывается уравнением:

$$y = bm^x.$$

Напомним, что значения экспоненциального тренда можно предсказывать при помощи функции **РОСТ** (подробно см. раздел 2).

Значения параметров экспоненциальной модели определяются при помощи функции **ЛГРФПРИБЛ**.

Линейный и экспоненциальный тренды тесно связаны между собой. Покажем это на рассматриваемом примере с продажами автомобилей.

Первоначально на основе имеющихся статистических данных по объему продаж за первые шесть недель сделаем прогноз на основе линейной модели ожидаемых продаж за последующие три недели (рис. 3.28). С этой целью введем в ячейки диапазона B8:B10 следующую формулу массива (ее ввод необходимо завершить нажатием комбинации клавиш <Shift>+<Ctrl>+<Enter>):

{= ТЕНДЕНЦИЯ (B2:B7; A2:A7; A2:A10)}

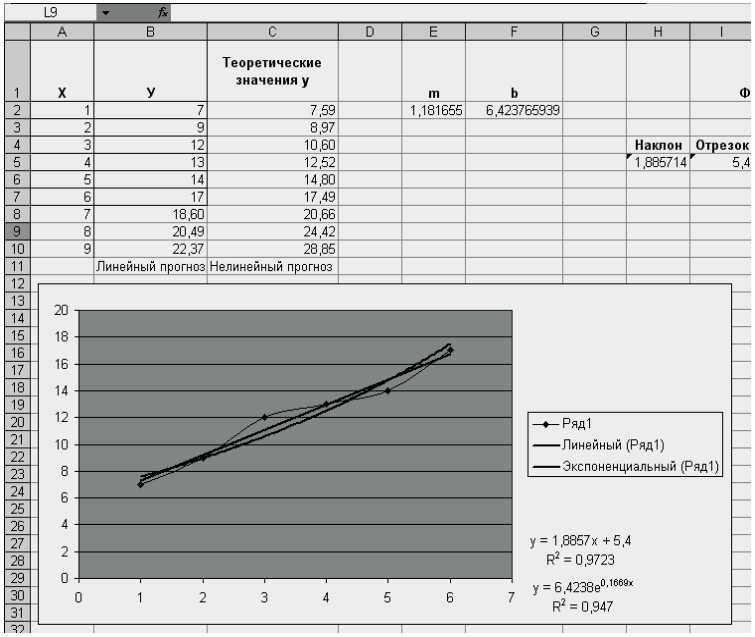


Рис 3.28. Связь между линейной и экспоненциальной линиями тренда

В диапазоне C2:C10 произведем прогноз на основе экспоненциальной модели. С это целью в ячейки этого диапазона введем следующую формулу массива (ее ввод необходимо завершить нажатием комбинации клавиш <Shift>+<Ctrl>+<Enter>):

{= РОСТ (B2:B7; A2:A7; A8:A10)}

Для определения параметров экспоненциальной модели в ячейке диапазона E2:F2 введем формулу массива

{= ЛГРФПРИБЛ (B2:B7; A2:A7)}

Квадрат коэффициента корреляции экспоненциальной модели равен 0,947 и меньше квадрата коэффициента корреляции линейной модели. Таким образом, в данном случае линейная модель более достоверно описывает зависимость между наблюдаемыми величинами.

ЗАДАНИЯ ДЛЯ САМОСТОЯТЕЛЬНОЙ РАБОТЫ

I. Выполните упражнения 1-26

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, 88 и 97 (используйте границы интервалов 70, 79, 89).

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. 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

7. 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.

8. 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?

9. 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:

x_1	x_2	x_3	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.

10. 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.

11. В рабочей зоне производились замеры концентрации вредного вещества. Получен ряд значений (в мг/м^3): 12, 16, 15, 14, 10, 20, 16, 14, 18, 14, 15, 17, 23, 16. Необходимо определить основные выборочные характеристики.

12. 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.

13. 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.

14. 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, 27, 27.

15. 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
12,6	12,8
12,3	13,2
11,9	13,0
12,2	12,9
13,0	13,5
12,4	13,1

16. 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?

17. 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?

18. 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

19. 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	82	79	91	77	68	86	81	90	85	94

Произошли ли статистически значимые изменения скорости чтения у студентов?

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

Образование	Зарплата сотрудника					
высшее	3200	3000	2600	2000	1900	1900
среднее спец.	2600	2000	2000	1900	1800	1700
среднее	2000	2000	1900	1800	1700	1700

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

Годы	Рождаемость	Смертность
1991	9,3	12,5
1992	7,4	13,5
1993	6,6	17,4
1994	7,1	17,2
1995	7,0	15,9
1996	6,6	14,2

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

Уровень инфляции	Ставка рефинансирования	Курс $
84	85	6,3
45	55	14
56	65	20
34	40	28
23	28	29

23. Построить зависимость зарплаты (р.) от возраста сотрудника гостиницы по следующим данным:

Возраст	20	50	45	40	25	30
Зарплата	800	2500	2500	2000	1200	1800

24. Построить зависимость жизненной емкости легких в литрах (Y) от роста в метрах (X_1) и возраста в годах (X_2) для группы мужчин:

X_1	X_2	Y
1,85	18	5,4
1,8	25	65,7
1,75	20	4,8
1,7	24	5,1
1,68	21	4,5
1,73	19	4,8
1,77	22	5,11
1,81	23	5,6
1,76	18	4,7

25. Определить должное значение жизненной емкости легких для мужчины возраста 22-х лет и роста 183 см из регрессионного уравнения, полученного в предыдущем упражнении.

26. Имеются данные о цене на нефть x (ден. ед.) и индексе акций нефтяных компаний y (усл. ед.):

x	17,28	17,05	18,30	18,80	19,20	18,50
y	537	534	550	555	560	552

Построить зависимость индекса акций нефтяных компаний от цены на нефть.

II. Уравнение регрессии

Построить линейную модель для двух наблюдаемых величин (например, объем реализованных подержанных автомобилей фирмой за указанное число недель) в соответствии с номером варианта.

Вариант 1

Неделя	1	2	3	4	5	6	7	8
Количество машин	9	15	24	29	38	46	52	58

Вариант 2

Неделя	1	2	3	4	5	6	7	8	9	10	11
Количество машин	15	22	26	33	40	45	51	58	63	69	78

Вариант 3

Неделя	1	2	3	4	5	6	7	8	9
Количество машин	14	23	30	39	45	54	63	70	78

Вариант 4

Неделя	1	2	3	4	5	6	7	8
Количество машин	12	18	25	32	40	46	53	60

Вариант 5

Неделя	1	2	3	4	5	6	7	8	9
Количество машин	10	18	22	28	34	39	46	51	54

Вариант 6

Неделя	1	2	3	4	5	6	7	8	9	10	11
Количество машин	12	17	23	30	35	40	48	54	59	65	72

Вариант 7

Неделя	1	2	3	4	5	6	7	8	9
Количество машин	12	21	30	36	44	54	61	70	78

Вариант 8

Неделя	1	2	3	4	5	6	7	8	9
Количество машин	7	17	19	28	35	42	41	52	57

Вариант 9

Неделя	1	2	3	4	5	6	7	8	9	10
Количество машин	9	16	20	27	34	39	44	52	58	64

Вариант 10

Неделя	1	2	3	4	5	6	7	8
Количество машин	13	19	26	30	37	44	49	55

ЛИТЕРАТУРА

1. *Пасько, В.* MicroSoft Office 2000 / В. Пасько. – К.: Изд. группа BHV, 2000.

2. *Ключников, М.В.* Применение MS Word и Excel в финансовых расче-тах: учеб. пособие / М.В. Ключников – М.: Market DS, 2006.

3. *Никольская, Ю.П.* Excel в помощь бухгалтеру и экономисту / Ю.П. Никольская, А. Спиридонов. – М.: Вершина, 2006.

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

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

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

7. 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.