Министерство образования и науки Российской Федерации
Уральский федеральный университет
имени первого Президента России Б.Н. Ельцина

В. Р. БАРАЗ, В. Ф. ПЕГАШКИН

ИСПОЛЬЗОВАНИЕ MS EXCEL ДЛЯ АНАЛИЗА СТАТИСТИЧЕСКИХ ДАННЫХ

Рекомендовано методическим советом ГОУ ВПО УрФу
в качестве учебного пособия для студентов,
обучающихся по направлениям подготовки
100700 – Коммерция (торговое дело),
150100 – Материаловедение и технология материалов
150400 – Металлургия

2-е издание, переработанное и дополненное

Нижний Тагил
2014

УДК 331.16:004.67
ББК С6:004.67
Б24

Р е ц е н з е н т ы:

кафедра технологии металлов Уральского государственного лесотехнического университета (зав. кафедрой профессор, д-р техн. наук Б.А. Потехин);
зав. лабораторией вычислительной техники Института материаловедения и металлургии Уральского федерального университета М.М. Розенбаум

Научный редактор доцент, канд. техн. наук С.И. Паршаков

Бараз, В.Р.

Б24 Использование MS Excel для анализа статистических данных : учеб. пособие / В. Р. Бараз, В. Ф. Пегашкин; М-во образования и науки РФ; ФГАОУ ВПО «УрФУ им. первого Президента России Б.Н.Ельцина», Нижнетагил. техн. ин-т (филиал). − 2-е изд., перераб. и доп. − Нижний Тагил : НТИ (филиал) УрФУ, 2014. – 181 с.

Предназначено для ознакомления с теоретическими положениями и приобретения практических навыков при изучении курса "Статистика" с использованием программы MS Excel. Рассмотрено большое количество примеров по обработке статистической информации.

Рекомендовано для студентов, обучающихся по направлению 100700 – «Коммерция (торговое дело)», а также для студентов других экономических и технических специальностей, изучающих соответствующие разделы курсов «Статистика» и «Организация эксперимента».

Библиогр. 6. Рис. 71. Табл. 21. Прил. 7.

УДК 331.16:004.67
ББК С6:004.67

ОГЛАВЛЕНИЕ

ВВЕДЕНИЕ

Статистика – в высшей мере логичный
и точный метод, позволяющий весьма
уклончиво формулировать полуправду
Из постулатов НАСА

Если мой сосед бьет жену каждый день,
а я никогда, то с точки зрения статисти-
ки мы оба бьем своих жен через день
Бернард Шоу

Статистика (немец. Statistik, от латинского status – состояние) рассматривается как наука *о методах изучения массовых явлений*. Некоторые процессы, наблюдаемые в *массовом* количестве, обнаруживают определенные *закономерности*, которые, однако невозможно заметить в отдельном случае или же при небольшом числе наблюдений.

Можно дать и иную формулировку: *статистика* – это наука, занимающаяся сбором и анализом данных о событиях, носящих *массовый* характер. При этом под *данными* принято понимать *любой вид зарегистрированной информации*.

Явления, которые в случае событий массового характера отличаются определенной закономерностью, однако не обнаруживаются на основе единичного наблюдения, называются *массовыми явлениями*. Сама такая закономерность называется *статистической закономерностью.*

Статистическая закономерность возникает в тех случаях, когда *а*) в исследуемом процессе действует *один общий* комплекс причин и когда *б*) наряду с этим в каждом отдельном случае действуют особые *дополнительные* причины, всякий раз иные.

При этом сами причины, которые определяют массовые процессы, принято делить на две категории:

– *основные причины*, которые действуют *во всех* случаях;

– *побочные (вторичные) причины*, которые проявляются только в *отдельных* случаях.

Скажем, возрастное старение человека определяется его биологической конституцией, социальными условиями. Все это, конечно, отражается на продолжительности жизни. Понятно, что названные факторы создают комплекс основных причин. Однако мы понимаем, что в жизни конкретного

человека добавляется множество дополнительных частных причин (неожиданная болезнь, стрессы, несчастный случай и проч.), которые порой самым прискорбным образом могут повлиять на его фактическую продолжительность жизни.

Если бы имели место только основные причины, то закономерность была бы абсолютной (т.е. для каждого элемента статистического массива одинаковой). Тогда её можно было бы уловить в каждом отдельном случае. Так, все люди жили бы одинаковое число лет. Вместе с тем если бы действовали только второстепенные причины, отличные для каждого случая, то никакой закономерности не было бы и воцарился бы полный хаос.

Таким образом, статистическая закономерность имеет место тогда, когда существует сочетание основных и побочных причин.

При этом можно добавить, что основные причины обусловливают само *существование* такой закономерности, а побочные причины определяют её *приблизительность*. Иначе говоря, закономерность проявляется только в массе случаев, а отдельный случай может отклоняться от общей картины. Можно полагать, что закономерность, вытекающая из постоянного действия основных причин, пробивается сквозь действие разнородных побочных факторов.

Из сказанного становится понятным, что статистика оказывается полезной в тех случаях, когда приходится анализировать процессы, которые при массовом наблюдении способны проявлять очевидную закономерность. Если бы действовали только главные причины, т.е. без наложения второстепенных, то все отдельные случаи происходили бы совершенно одинаково и не было бы нужды анализировать всю их массу. Достаточно было бы взять один из случаев и на его основе сделать выводы, относящиеся уже ко всей исследуемой совокупности.

Так, кстати сказать, поступают во многих науках. Например, в химии полагают, что капля воды похожа на другую. Проводят анализ одной пробы воды и на его основе делают обобщение относительно химического состава воды. Похожим образом примерно действуют в биологии или анатомии. Например, исследуется анатомическое строение одной собаки, и делаются выводы об анатомическом строении всех собак.

Там же, где закономерность пробивается через результаты воздействия побочных причин, приходится изучать уже целую массу случаев, чтобы иметь возможность выявить закономерность. В такой ситуации исследование единичного примера может привести к ложным заключениям.

В массовых процессах обычно различают два элемента: *систематический* (*постоянный*) и *случайный* (*побочный*). Систематический элемент явля-

ется результатом действия *основных* причин, случайный элемент – это следствие действия *побочных* причин (действуют по-разному в каждом отдельном случае).

Статистическая закономерность проявляться более отчетливо в случае действия *закона больших чисел*. Этот закон отражает закономерности, присущие случайным *событиям массового* характера. При большом количестве наблюдений влияние *случайных* факторов *взаимно уравновешивается* и вступают в действие *главные причины*, которые отражаются в некотором *постоянстве средних* чисел.

Например, каждый покупатель в магазине выбирает именно тот товар, который в данный момент ему нужен. Но в целом по магазину можно сравнительно точно предвидеть общий объем спроса, его структуру за год, отдельные сезоны и даже дни недели. Для выявления конкретных закономерностей покупательского спроса и нужна статистическая информация, отображающая специфику спроса по дням недели, времени года и в целом за год.

Для выполнения закона больших чисел важно соблюсти определенные условия.

1. Исследуемый *массив* должен быть *однородным*, т.е. быть одинакового качества. Это означает, что *все элементы массива* подпадают под действие *одних и тех же основных причин*. В противном случае могут возникнуть иные основные факторы и тогда общую картину выявить окажется невозможным.

Однородна ли данная статистическая масса – этого нельзя установить на основе статистического исследования. Для этого нужен качественный анализ, который проводится методами, применяемыми в соответствующих областях науки (физические, экономические и т.д.).

2. Побочные причины, воздействующие на разные элементы массива, должны быть *независимыми друг от друга* или же мало зависимыми.

Таким образом, не может быть хорошей статистики там, где нет достаточно *а*) *многочисленных*, *б*) *однородных* и *в*) *независимых* данных. Если это условие не соблюдено, то отсутствует и подлинная статистика.

В курсе общей теории статистики принято условно различать описательную статистику и аналитическую. *Описательная статистика* преимущественно связана с планированием исследования, сбором информации и представлением полученных результатов в виде статистических показателей. Удобная форма представления статистической информации – таблицы, графики. Задача *аналитической статистики* – выявить причинные связи, оценить влияние исследуемых факторов и сделать надлежащие выводы, на ос-

новании которых могут быть приняты ответственные решения. Часто исследуемый процесс представляется в аналитической форме, т.е. в виде уравнения (эмпирической формулы).

Знание статистики помогает нам принять хорошие, лучше сказать, оптимальные решения. При этом статистика отнюдь не отвергает опыт и интуицию. Её можно рассматривать как один из компонентов процесса принятия решения, но отнюдь не весь процесс. Поэтому оправданно полагать, что статистика дополняет, но не заменяет деловой опыт, здравый смысл и интуицию.

И, наконец, не следует забывать – использование статистики становится все более важным преимуществом в конкуренции.

Мощным инструментальным средством при выполнении статистических исследований является использование компьютерной техники. В этой связи широкое распространение в деловой сфере (понимай – в коммерческой деятельности) получили специальные пакеты прикладных программ. Они позволяют обеспечить весьма впечатляющую быстроту статистических расчетов, высокую надежность и достоверность результатов, возможность легко представлять данные в аналитической, графической или табличной формах.

Среди подобных программ большой известностью пользуется приложение Microsoft Excel, которое включает в себя программную надстройку "Пакет анализа" и богатую библиотеку из большого числа статистических функций.

Основное назначение данного учебного пособия – познакомить студентов с поразительными возможностями этого весьма полезного приложения и показать, как его удобно применять для выполнения достаточно стандартных статистических расчетов применительно к работе в деловой сфере. Таким образом, оно адресовано, прежде всего, студентам, обучающимся по специальностям "Коммерция (торговое дело)", "Мировой рынок сырья и металлов". Вместе с тем методический способ изложения материала, приводимые практические примеры носят достаточно общий характер. Поэтому данное пособие может оказаться пригодным для студентов и других специальностей, изучающих в соответствующих учебных дисциплинах методы статистического анализа данных.

Первое издание учебного пособия "Использование MS Excel для анализа статистических данных" было выпущено в свет в 2007 году. С той поры накоплен определенный опыт, позволяющий на основании мнения студенческой аудитории, а также университетских коллег судить о пригодности и полезности данного пособия в качестве учебного материала при изучении приемов статистического исследования. При этом авторы сочли возможным ввe-

сти некоторые добавления (сведения о методах измерения, компьютерный расчет уравнений регрессии), а также сделаны некоторые коррективы в обсуждаемых примерах, касающихся конкретного применения Excel.

Основное содержание данного пособия состоит не только в развернутом изложении известных методов статистического исследования, но и в описании приемов применения в практике коммерческой деятельности выборочного метода, корреляционно-регрессионного анализа, а также динамических рядов и способов перспективного прогнозирования.

Каждая глава пособия условно поделена на две части. Первая часть содержит изложение основных положений, касающихся рассмотрения соответствующего раздела теории статистики. Вторая часть главы – это практикум, где мы, что называется, засучив рукава, уже на деле применяем усвоенные теоретические положения, используя впечатляющие возможности компьютерной программы Excel.

Следует сделать одно замечание. Данное пособие вовсе не претендует на подробное и последовательное изложение всех положений общей теории статистики. Предполагается, что студент знаком с содержательной стороной этой дисциплины и вполне владеет каноническими сведениями из теории статистики (методы группировок, абсолютные и относительные величины, средние величины, показатели вариации, ряды распределения и ряды динамики, измерение связи и т.д.). Поэтому если в пособии и приводятся подробные описания некоторых теоретических положений, то они излагаются для того, чтобы дать более понятное объяснение использования компьютерной технологии при статистическом анализе сугубо практических ситуаций.

Предложенные для рассмотрения примеры по своему содержанию намеренно носят довольно иронично-шутливый характер. Поэтому избыточно серьезный читатель, а тем более достаточно въедливый, легко найдет в этом очевидные изъяны. Однако использование такого методического подхода преследовало вполне понятную цель – в легкой и по возможности непринужденной манере попытаться рассказать о вещах, в общем-то, довольно скучных, если не сказать просто занудных, однако не теряющих от этого свою несомненную важность и очевидную полезность.

1. ВЫБОРОЧНЫЙ МЕТОД СТАТИСТИЧЕСКОГО АНАЛИЗА

Работая над решением задачи, всегда полезно знать ответ.

Закон Мэрфи

Коля ловил девчонок, окунал их в лужу и старательно измерял глубину погружения каждой девчонки, а Толя стоял рядышком и смотрел, как девчонки барахтаются. Чем отличаются Колины действия от Толиных и как такие действия называют физики?

И физики и химики назовут Колины и Толины действия хулиганством и надают им по шее. Однако с точки зрения бесстрастной науки Толя производил наблюдение, а Коля ставил опыты.

Григорий Остер. "Сборник задач по физике"

Как было отмечено, цель статистического исследования состоит в отыскании определенных закономерностей в событиях массового характера, каждое из которых по отдельности имеет достаточно случайное проявление. Для достижения этой цели используются специальные статистические приемы, основанные на реализации так называемого выборочного изучения. Познакомимся с основными его положениями.

1.1. Измерение

Предварительно напомним вполне очевидные сведения. Известно, что любое статистическое исследование включает обязательную процедуру проведения измерений. Само *измерение* определяется как способ нахождения значения физической величины опытным путем с помощью специальных технических средств. Сущность измерения фактически состоит в сравнении двух физических величин – измеряемой и известной. Первая отражает особенность исследуемого объекта (например, в коммерции это может быть количество реализованного товара в физическом объеме или в стоимостном

выражении), вторая присуща специально созданному объекту – эталону или мере. Сравнение этих объектов сводится к сопоставлению их размеров, следовательно, основывается на выявлении их количественного соотношения. При этом сравниваемые величины должны быть однородными, т.е. имеющими сходную физическую природу, одинаковую размерность.

В качестве эталона могут применяться самые различные меры, порой весьма своеобразные. Так, в старину в качестве меры для определения расстояния служил локоть (это примерно 0,5 м, т.е. длина локтевой кости человека). Можно привести и просто забавный пример. В известном детском мультфильме "38 попугаев" Мартышка, Слоненок и Попугай измеряли длину Удава. А в качестве эталона использовали самих себя, что позволило потом Удаву горделиво заключить: "А в попугаях-то я длиннее!" Словом, говоря философски, можно измерять в чем угодно, было бы только что измерять.

При экспериментальном определении какой-либо величины приходится сталкиваться с тем, что параллельные измерения не дают одинаковых результатов даже при самой тщательной подготовке опыта. Это обстоятельство является следствием того, что на процесс измерения и, стало быть, на его результат оказывает влияние огромное число факторов (начиная от погоды, температуры, степени изношенности оборудования, измерительного инструмента и кончая эмоциональным состоянием экспериментатора в момент измерения).

Влияние каждого фактора в отдельности может быть в целом совершенно ничтожным, но в совокупности они способны вызывать случайные (и потому непредсказуемые) отклонения измеряемой величины от ее истинного значения. Это означает, что при проведении повторных измерений одной и той же физической величины мы получим в итоге несколько отличающиеся друг от друга результаты.

Таким образом, измеренное значение определяется, с одной стороны, влиянием основных факторов, а с другой, параметрами, обусловленными случайными причинами.

Измерения принято делить на прямые и косвенные. Основным признаком является вид уравнения измерения, связывающее измеряемую (искомую) величину и непосредственно наблюдаемую (эталон).

Прямые измерения – измеряемая величина A пропорциональна непосредственно наблюдаемой B, т.е. получается непосредственно с помощью измерительного прибора (используется непосредственный счет единиц наблюдения).

Аналитически такое измерение можно представить в виде линейного соотношения $A = \kappa B$, где κ – заданный коэффициент.

В качестве примера можно указать измерение массы на циферблатных или равноплечных весах или измерение температуры термометром. Получаемые данные – это абсолютные значения.

Косвенные измерения – в этом случае измеряемая величина A является известной функцией непосредственно наблюдаемого аргумента B и определяется в результате математических действий над результатами прямых измерений. Это значит, что на основании результатов изучения одного процесса с использованием известной аналитической зависимости (уравнения) получаются сведения о другом. Типичный пример: измерение плотности твердого тела по его массе и геометрическим размерам.

Указанное соотношение имеет следующий вид: $A=f(B)$; получаемые данные являются относительными величинами.

Обычно удается провести ограниченное число параллельных измерений или получить случайную *выборку* (т.е. конкретный набор экспериментальных данных) из *генеральной совокупности* (всё мыслимое количество повторных измерений)*. В этом случае задача исследователя состоит в том, чтобы по такой выборке (т.е. на основе знания части целого) получить истинное значение (или так называемое математическое ожидание) самого целого (генеральной совокупности).

В связи с этим задача статистической обработки сводится к следующему:

1. *Отыскать истинное значение* измеряемой величины $\tilde{x}$, однако в большинстве случаев оно оказывается неизвестным. Поэтому его заменяют некоторым приближенным значением, которое наиболее вероятно соответствует истинному значению. В статистике показано, что этому условию наиболее полно отвечает среднее арифметическое $\bar{x}$ выборочной совокупности.

2. *Оценить погрешность* (ошибку) Δx, с которой найдена эта истинная величина; иными словами, нужно определить ту величину, на которую отличается приближенное значение $\bar{x}$ от истинного $\tilde{x}$.

Типы ошибок измерения и их определение. Итак, погрешность (ошибка) измерения – это отклонение результатов измерения от истинного значения измеряемой величины.

* Подробнее см. в разделе 1.3.

По форме представления (по отношению к измеряемой величине) различают следующие ошибки**.**

1. *Абсолютная ошибка* – разность между измеряемым (приближенным) значением $x_{изм}$ и истинной величиной $x_{ист}$:

$$\Delta x = x_{изм} - x_{ист} .$$

Здесь надо дать некоторое пояснение. В общем случае само измеренное значение можно записать как x_i , в качестве истинного значения принято указывать его приближенное значение в виде среднего арифметического $\bar{x}$, поскольку собственно истинное значение $\tilde{x}$, как отмечалось, обычно остается неизвестным. Поэтому величину абсолютной ошибки принято записывать в виде выражения

$$\Delta x = \bar{x} - x_i .$$

2. *Относительная ошибка* – погрешность измерения, выраженная отношением абсолютной погрешности измерения к истинному значению (точнее, к его приближенному значению):

$$\delta = \Delta x / \bar{x}$$

Относительная погрешность является безразмерной величиной, либо измеряется в процентах.

Абсолютная ошибка характеризует погрешность метода, который был выбран для измерения. Относительная ошибка характеризует качество измерений. Точностью измерения называют величину, обратную относительной ошибке, т.е. 1/δ.

По характеру проявления – различают систематические, случайные и грубые погрешности измерения.

1. *Систематические ошибки* – порождается причинами, действующими а) регулярно и б) в определенном направлении. Они могут быть связаны с ошибками приборов (неправильная шкала, калибровка и т. п.), неучтёнными экспериментатором; с индивидуальными ошибками экспериментатора; ошибками метода анализа.

Исключаются путем введения поправок, найденных экспериментальным путем (например, градуировка термопар).

2. *Случайные ошибки* – составляющие погрешности измерения, изменяющиеся случайным образом в серии повторных измерений одной и той же величины, проведенных в одних и тех же условиях. В появлении таких ошибок не наблюдается какой-либо закономерности, они обнаруживаются при повторных измерениях одной и той же величины в виде некоторого разброса получаемых результатов. Случайные ошибки неизбежны, неустранимы и всегда присутствуют в результате измерения, однако их влияние, как правило, можно устранить статистической обработкой.

3. *Грубые ошибки (промахи или выскакивающие значения)* – погрешности, возникшие вследствие недосмотра экспериментатора или неисправности измерительной аппаратуры (например, если это неправильно зафиксированный номер деления на шкале прибора или ошибочно выполненный отсчета, ненадлежащая юстировка прибора и т.п.).

В зависимости от характеристик измеряемой величины для определения ошибок измерений используют различные методы.

1. *Метод Корнфельда* – заключается в выборе интервала в пределах от минимального $x_{\min}$ до максимального $x_{\max}$ результата измерений и погрешность рассчитывается как половина разности между этими крайними величинами измерения:

$$\Delta x = \frac{x_{\max} - x_{\min}}{2}$$

2. *Средняя квадратичная погрешность* (среднеквадратичное отклонение) S_n:

$$S_n = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}},$$

где x_i – измеренные значения элементов выборки, $\bar{x}$ – среднее арифметическое выборки и n – её размер.

3. *Средняя квадратичная погрешность среднего арифметического* (стандартная ошибка) $\boldsymbol{S}_{\bar{x}}$:

$$S_{\bar{x}} = S_n / \sqrt{n}.$$

В статистике для оценки погрешности наиболее часто используемым показателем является среднеквадратичное отклонение (СКО) [*].

1.2. Понятие о выборном исследовании

Представим себе следующую ситуацию.

Вы являетесь руководителем аналитического отдела фирмы, занимающейся поставками изделий прокатного производства для строительно-монтажных организаций. И вот однажды, еще не успев выпить чашку утреннего кофе, вы получаете по телефону строгое указание шефа: к завтрашнему совещанию подготовить материалы о реакции клиентов фирмы относительно планируемого изменения сроков поставок и размера ценовых скидок.

Как надлежит поступить в этом случае? Ситуация достаточно типичная и некоторые организационные мероприятия представляются вполне очевидными. Понятно, что следует выслушать мнение клиентов, для чего необходимо с ними переговорить. Вы начинаете размышлять. Конечно, идеальный случай – это попытаться обзвонить всех клиентов. Однако такой вариант решительно нереален. Во-первых, число клиентов фирмы весьма приличное – свыше пяти сотен; следовательно, если на каждый разговор потратить что-то около 10 минут, то на всё потребуется уйма времени, а также случатся избыточные траты на телефонные переговоры (разбирайся потом с бухгалтерией!). Во-вторых, у вас небольшой штат сотрудников и не всем можно было поручить такую деликатную миссию. Наконец, сроки – уже к завтрашнему утру на столе шефа должна лежать служебная записка с требуемыми материалами.

Итак, что делать? Выход очевиден: нужно построить *выборку* из *генеральной совокупности* всех клиентов фирмы, внесенных в базу данных. В этом случае придется обзвонить лишь ограниченное количество клиентов. И на основании этой информации постараться отразить вероятную реакцию *всех* потребителей, а не только тех, кто попадут в вашу небольшую выборку. Таким образом, на основе изучения *части* постараться получить достоверное представление о *целом*. В этом и состоит идея метода выборочного исследования.

В теории статистики даются четкие рекомендации относительно того, как на основании фактического знания о малом получить надежную информацию о неизвестном многом. Так, например, сколько нужно исследовать единиц наблюдения (каков должен быть объем выборки), каким образом ор-

[*] Более детально об этом изложено в разделе 1.5.

ганизовать отбор, какие, наконец, нужно рассчитать показатели, которые дадут надежное представление об изучаемом процессе в целом.

1.3. Основные определения

Выборочный метод – это такой способ статистического исследования, при котором *обобщающие показатели* изучаемого *массива* устанавливаются по *некоторой его части* на основе положений *случайного отбора*. При указанном методе обследованию подвергается сравнительно небольшая часть изучаемой совокупности. При этом подлежащая изучению статистическая совокупность, из которой проводится отбор части единиц, называется *генеральной совокупностью*. Иначе говоря, генеральная совокупность – это набор элементов (люди, объекты и проч.), которые нужно изучить.

Отобранная из генеральной совокупности некоторая часть единиц, подвергающаяся обследованию, называется *выборочной совокупностью* или *выборкой*. Следовательно, выборка – это небольшой набор объектов, извлеченных из генеральной совокупности.

Простейший пример выборочного метода – экзамен, когда нужно проверить знания студента. Однако спрашивать студента обо всем невозможно, поэтому необходимо сделать выборку из его знаний. Сделать это можно двумя способами:

– вопросы записываются на билеты, билеты перемешиваются, а студент сам выбирает их в случайном порядке;

– профессор решает, о чем спросить студента, и направленно отбирает вопросы таким образом, чтобы охватить ими всю программу или какую-то тему по данному курсу.

Понятно, что изучаемая совокупность – это сумма знаний, которыми должен обладать студент, а задаваемые вопросы – выборочная совокупность.

Точность результатов выборочных обследований много раз проверялась. Эти наблюдения подтвердили, что результаты таких обследований дают точное представление об изучаемой совокупности и могут применяться на практике без опасений серьезных ошибок, если выборка действительно репрезентативна и ее численность установлена на основе теории вероятностей.

Заметим, что из генеральной совокупности можно отобрать огромное число выборок. Например, при генеральной совокупности $\boldsymbol{N}$, равной 100 элементам, можно извлечь выборки объемом $\boldsymbol{n}$ =10 в количестве $17 \cdot 10^{12}$ вариантов (!).

Ценность выборочного обследования состоит в следующем:

– требует меньших затрат, чем сплошное наблюдение (т.е. изучение всей совокупности);

– позволяет значительно раньше получать результаты статистического исследования (порой это может быть решающим фактором);

– в ряде случаев может быть только единственным способом, если обследование сопровождается разрушением или уничтожением элемента совокупности (прочность тканей на разрыв, механические испытания металлический изделий, установление носкости обуви и проч.).

Важная особенность – в основе отбора единиц для обследования положен принцип равных возможностей попадания в выборку каждой единицы генеральной совокупности.

Такой подход позволяет:

– *исключить* формирование выборочной совокупности за счет *лучших* или *худших* образцов;

– *предупредить* появление *систематических* (тенденциозных) ошибок;

– *дать* количественную *оценку ошибки представительства (репрезентативности).*

1.4. Репрезентативность выборки

Процесс построения выборки – из большей по размеру генеральной совокупности извлекается выборка для проведения измерений и подробного анализа. При этом полагается, что выборка является *репрезентативной* (представительной).

Суть репрезентативности выборки – выборка (часть целого) должна достоверно отражать генеральную совокупность (само целое). Этому соответствует одинаковость частот проявления признака (свойства) как для выборки, так и для всей совокупности, т.е. кривые распределения должны быть идентичными (положение центра, характер формы кривой). Различие только по размаху вариации (дисперсии) – генеральная совокупность должна иметь меньший разброс относительно среднего.

Таким образом, выборка – это результат непосредственного наблюдения части целого, позволяющее косвенно судить о самом целом.

Для обеспечения репрезентативности выборки применяются два метода:

– *отбор в случайном порядке*, при таком отборе каждый элемент совокупности имеет одинаковый шанс попасть в выборку;

– *направленный отбор*, в этом случае отбираются только некоторые единицы (на основе выработанных специальных критериев).

Для практической работы нужно иметь *основу генеральной совокупности*, которая даст возможность обращаться к отдельным элементам по номерам. Так, основа может иметь вид списка объектов генеральной совокупности, которым присвоены номера от 1 до N, где N – число объектов (объем) генеральной совокупности.

Следовательно, основа совокупности (нумерованный список) позволяет из имеющегося объема в интервале от 1 до N получить доступ к элементам генеральной совокупности и сформировать из них выборку объемом n.

Выборка, которая включает полную генеральную совокупность, называется *переписью*.

Принято различать *два типа выборки*. После того как объект извлечен из генеральной совокупности для включения его в выборку, его *а*) либо *возвращают обратно* в генеральную совокупности (тогда он может попасть в эту выборку повторно), *б*) либо он *не возвращается*.

Соответственно бывают следующие две комбинации:

а) *выборка с возвратом* или *повторный отбор* (объект генеральной совокупности может попасть в выборку более одного раза) и

б) *выборка без возврата* или *бесповторный отбор* (при этом все объекты выборки получаются разными).

Отметим принятые определения:

1. *Параметр выборки* (или *выборочный параметр*) – показатель, вычисленный на основе данных выборки (т.е. это любое число, рассчитанное из данных выборки). Например, таким параметром является среднее арифметическое выборки $\bar{x}$, стандартное отклонение S_n.

2. *Параметр генеральной совокупности* (или просто *параметр*) – это показатель (число), рассчитанный для всей этой совокупности. Пример – среднее арифметическое (истинное значение) $\tilde{x}$ и стандартное отклонение самой генеральной совокупности σ. Параметр является *фиксированным* числом, т.к. при его вычислении отсутствует случайность. Однако обычно мы не знаем всех данных по генеральной совокупности, поэтому параметр является *неизвестной* величиной.

Обычно существует соответствие между параметром выборки и параметром генеральной совокупности. Для каждого параметра генеральной совокупности (его значение, повторим, мы не знаем) существует выборочный

параметр, рассчитанный на основе данных, представляющих наилучшую доступную информацию о неизвестном параметре генеральной совокупности. Такой выборочный параметр называют *оценочной функцией* параметра генеральной совокупности, а его фактическое значение, вычисленное из элементов выборки, называют *оценкой* параметра совокупности. Например, среднее арифметическое выборки является оценочной функцией среднего арифметического совокупности. Иными словами, среднее арифметическое выборки позволяет *приближенно* судить о среднем арифметическом совокупности (просто лучшей информации об этом неизвестном у нас нет).

1.5. О выборочном распределении

Итак, любая выборочная совокупность, полученная на основе случайного отбора изучаемых элементов из генеральной совокупности, позволяет перейти от информации о выборке (которая у нас есть) к информации о генеральной совокупности (которую нам хотелось бы иметь). Понятно, что можно организовать повторную выборку и для неё получить свое среднее арифметическое, которое тоже отражает поведение генеральной совокупности. А затем третье, четвертое и т.д. Иными словами, можно получить набор из нескольких выборок, извлеченных из одной и той же генеральной совокупности. В результате мы получим еще один статистический массив, например, из средних арифметических этого набора выборок (так называемое *выборочное распределение*). И для этого массива можно рассчитать свое среднее, которое более надежно будет характеризовать (приближенно, конечно, но с меньшей погрешностью) интересующую нас генеральную совокупность.

Пояснение. Положим, мы готовим отчет о результатах опроса случайно отобранных покупателей и видим, что они тратят на покупку кондитерских товаров в среднем 112,6 руб. за один визит в магазин. Это число для нас не выглядит случайным. *Но как результат нашего обследования* оно является *случайным.*

Как это понять? Дело в том, что число 112,6 само по себе не есть случайность. Однако оно отражает "средние расходы на кондитерские товары за один визит для случайно выбранных покупателей". А это есть случайная переменная. Ее случайность объясняется тем, что в результате выполнения случайного эксперимента каждый раз опрашивается новая случайная выборка покупателей и, следовательно, каждый раз будет получен иной результат.

Здесь 112,6 руб. – *конкретный* результат измерения (как нечто среднее). Но если повторять такие наблюдения, допустим, каждый раз в течение десяти дней, то получим уже набор *случайных* величин (ибо заранее точно указать среднюю трату на покупку в данный день исследуемой декады невозможно). Следовательно, проведя в течение 10 дней такие наблюдения, будем иметь набор из 10-ти конкретных

средних. Можно оценить распределение этих средних (они же случайные числа и характеризуются изменчивостью) и рассчитать для них свое среднее и свое стандартное отклонение.

Итак, когда исследование проводится только один раз, полученные результаты – это просто конкретные числа. Однако нужно также понимать, что вследствие ограничений реальной жизни мы действительно проводим исследование, как правило, только один раз. *Предположение о многократном повторении* исследования – это *лишь способ понять имеющийся фактический результат*. Следовательно, на основе анализа *одной* выборки, одного значения параметра выборки (допустим, среднего арифметического или стандартного отклонения), мы пытаемся интерпретировать *все* остальные результаты, которые *могли бы иметь место*.

1.6. Стандартная ошибка как оценка стандартного отклонения

В реальной жизни обычно нет возможности работать непосредственно с выборочным распределением (у нас *нет* нескольких выборок). При этом нужно помнить, что его параметры определяются свойствами *всей* генеральной совокупности. Информацию же мы имеем только для *одной* выборки.

Если имеется набор из нескольких выборок, сформированный на базе одной и той же генеральной совокупности, то полученный комплект, например, средних арифметических этих выборок сформирует свой ранжированный массив (выборочное распределение этих средних). И для него можно рассчитать свое среднее арифметическое $\overline{X}$ (среднее средних).

При этом наблюдается вот какая интересная особенность – само выборочное распределение (средних набора выборок $\overline{x}_i$) близко к *нормальному* виду, хотя генеральная совокупность объектов может и отличаться от нормального распределения.

Пояснение. Напомним, что нормальным (симметричным) распределением называется такое, в котором частоты двух любых вариантов, равно отстоящих в обе стороны от центра распределения (среднего арифметического $\overline{x}$), равны между собой. Кривая нормального распределения по форме напоминает симметричный колокол. Для такого распределения имеет место равенство показателей центра распределения: среднего арифметического $\overline{x}$, моды *Mo* (наиболее часто повторяемого результата в данном массиве) и медианы *Me* (она делит ранжированный вариационный ряд на две равные части).

Для случая несимметричного (скошенного) распределения вводится понятие *асимметрии* кривой распределения (значения данных на одной стороне кривой затухают быстрее, чем на другой). Простейший показатель асимметрии основан на соот-

ношении значений центра распределения: *чем больше разница между средними ($\bar{x}$ – Мо), тем больше асимметрия ряда.*

Относительно асимметрии важно знать – многие статистические методы требуют, чтобы данные были (хотя бы приблизительно) нормально распределенными. Если эти методы применяются к несимметричным рядам, то полученный результат будет неточным или же просто ошибочным. И даже если результаты получаются в основном корректными, будет определенная потеря эффективности анализа, т.к. не обеспечивается наилучшее использование всей информации, содержащейся в наборе данных.

Выход – использование такого преобразования, которое переводит несимметричное распределение в более симметричное. Преобразование заключается в замене каждого значения набора данных другим числом (например, логарифм этого значения) с целью упростить статистический анализ.

Наиболее распространенный прием в статистике – это логарифмирование, которое можно использовать только для положительных чисел. Логарифмирование часто преобразует скошенные ряды в симметричные, поскольку происходит растягивание шкалы возле нуля.

Для *симметричных* распределений рассчитывается специальный показатель – *эксцесс*, характеризующий островершинность кривой. Фактически эксцесс представляет собой выпад вершины эмпирического распределения верх или вниз относительно вершины кривой нормального распределения.

В теории статистики показано, что в случае нормального распределения значения средних (для отдельной выборки $\bar{x}$ и набора выборок $\bar{X}$) остаются теми же, в то время как их вариации (например, в виде стандартных отклонений) различаются. Переход от единичной выборки к набору выборок (полученных извлечением из одной и той же генеральной совокупности) приводит к уменьшению изменчивости, что отражается в снижении величины стандартного отклонения.

В теории статистики доказывается, что стандартное отклонение *среднего выборочной совокупности* $\sigma_{\bar{x}}$ определяется по формуле

$$\sigma_{\bar{x}} = \sigma / \sqrt{n} \ ,$$

где σ – стандартное отклонение генеральной совокупности; n – объем выборки, извлеченной из генеральной совокупности объемом N.

Но поскольку мы работаем с выборкой, то значение σ (а это показатель генеральной совокупности) нам неизвестно. Следовательно, неизвестно и стандартное отклонение среднего выборочной совокупности. Однако, располагая фактическим набором единиц наблюдения, входящих в выборочный массив, можно рассчитать *стандартное отклонение выборки* S_n по формуле:

$$S_n = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n-1}} \ ,$$

где x_i – измеренные значения признака элементов выборки, $\bar{x}$ – среднее арифметическое выборки и n – её размер.

При этом стандартное отклонение генеральной совокупности принято определять по выражению

$$\sigma = \sqrt{\frac{\sum (x_i - \tilde{x})^2}{N}},$$

где $\tilde{x}$ – среднее арифметическое генеральной совокупности.

Размерность у стандартного отклонения та же, что и у исходных данных (рубли, количество автомашин, тонны и т.д.).

Теперь, зная стандартное отклонение S_n, можно вычислить *стандартную (случайную) ошибку* выборочного распределения $S_{\bar{x}}$ по формуле (для случая *повторного* отбора):

$$S_{\bar{x}} = S_n / \sqrt{n}.$$

Это можно прокомментировать следующим образом:

Стандартная (случайная) ошибка среднего арифметического равна стандартному отклонению отдельных результатов, деленному на корень квадратный из числа измерений.

Стандартная ошибка грубо показывает, насколько мы ошибались, используя лучшую доступную выборочную информацию (например, среднюю стоимость покупок 100 *случайных* покупателей) вместо недоступной информации о генеральной совокупности (средняя стоимость покупок *всех* покупателей города).

Пояснение. Так в чем же принципиальное различие между $S_{\bar{x}}$ и S_n?
Стандартное (*среднеквадратичное*) *отклонение* S_n приближенно показывает, насколько *отдельные значения элементов выборки* отличаются от *среднего значения набора данных* этой выборки (т.е. x_i от $\bar{x}$). *Стандартная* (*случайная*) *ошибка* $S_{\bar{x}}$ приближенно показывает, насколько *среднее* $\bar{x}$ отличается от *среднего значения генеральной совокупности (истинного значения)* $\tilde{x}$.

Пояснение. В расчетах, как было сказано, обычно пользуются результатами обследования одной выборки. Поэтому в качестве среднего фигурирует среднее арифметическое этой единственной выборки $\bar{x}$. И именно его сопоставляют со средним арифметическим генеральной совокупности (истинным значением) $\tilde{x}$. Это значит, что величина выборочного среднего $\bar{X}$ (для этого случая полагают, что было извлечено несколько выборок) фактически не используется, вместо него берут реальный результат расчета – среднее арифметическое единственной выборки $\bar{x}$. Этот

выбор основывается на утверждении, что при нормальном распределении значения этих средних совпадают, хотя они и различаются вариациями.

Следовательно, вычисление собственно ошибки выборки $S_{\bar{x}}$ (она свидетельствует о том, как различаются между собой $\tilde{x}$ и $\bar{x}$) ведется с использованием среднеквадратичной ошибки S_n, уменьшенной на величину $\sqrt{n}$.

Итак, резюме.

1. Выборочное распределение – это распределение, построенное на анализе средних арифметических нескольких (числом равных n) выборочных совокупностей, извлеченных из общей (генеральной) совокупности.

2. Если извлечена одна выборка, то строится распределение самих элементов совокупности, которые реально наблюдаемы:

а) определяется среднее арифметическое для этой выборки $\bar{x}$;

б) вычисляется стандартное отклонение (среднеквадратичное отклонение) S_n (при $n \to \infty$ имеем $lim\ S_n \to \sigma$).

3. Если извлечено n выборок, то для них самих:

а) вычисляются средние (для каждой из n выборок);

б) описывается распределение этих средних $\bar{x}_i$ (т.е. строится кривая распределения средних);

в) определяется среднее арифметическое (интегральное) для этого ряда конкретных средних $\bar{X}$;

г) вычисляется своя стандартная ошибка $S_{\bar{x}}$, которая является стандартным отклонением среднего арифметического $\bar{X}$ самого выборочного распределения. Она связано со стандартным отклонением S_n соотношением $S_{\bar{x}} = S_n / \sqrt{n}$.

4. Поскольку фактически приходится иметь дело с одной выборкой, то выборочное среднее $\bar{X}$ заменяется реальным показателем в виде среднего арифметического $\bar{x}$ этой выборки, которое затем и сопоставляется со средним генеральной совокупности $\tilde{x}$.

Сказанное можно проиллюстрировать следующей схемой (рис.1.1).

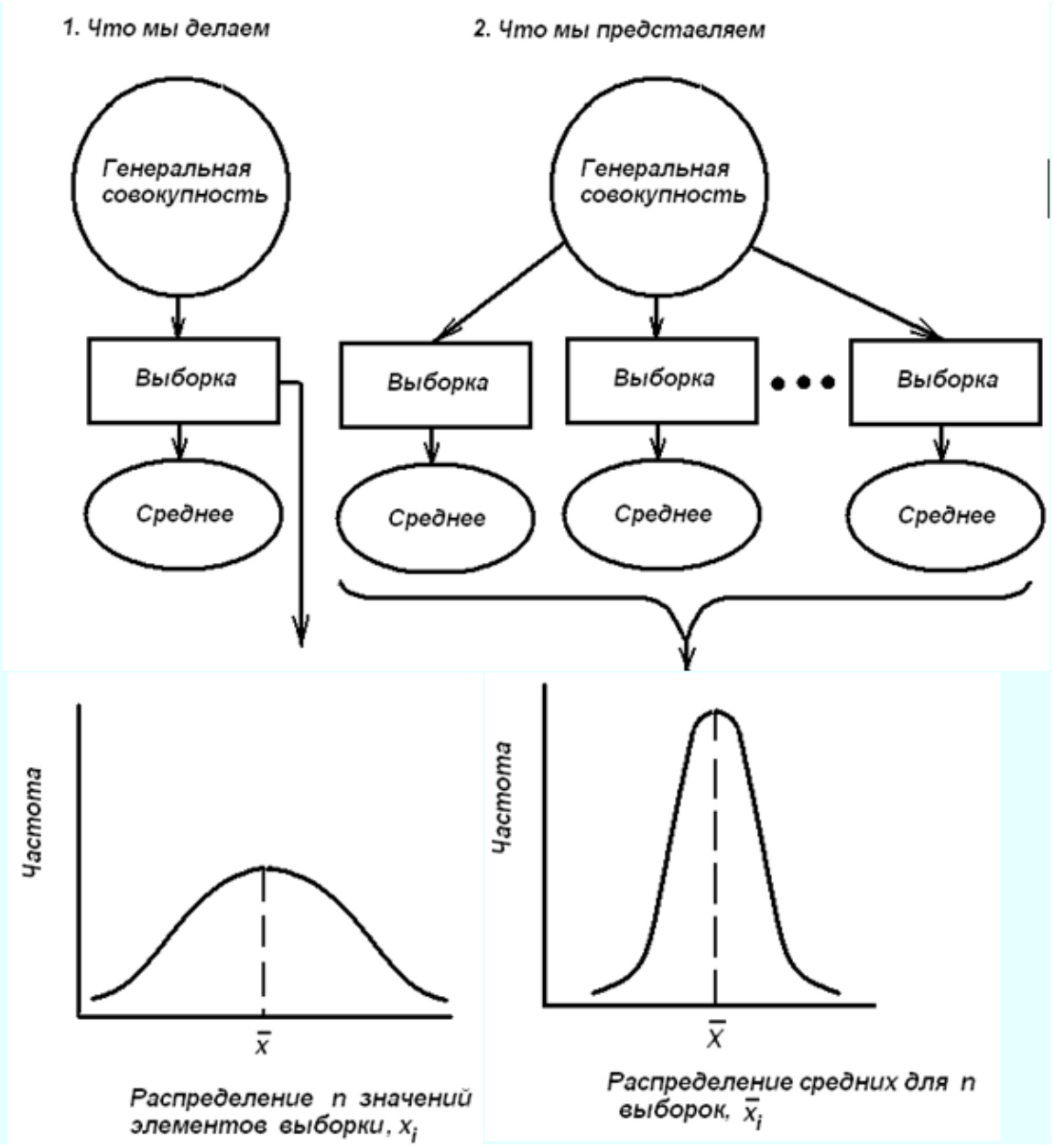


Рис.1.1. Фактический (1) и предполагаемый (2) набор и анализ выборок, сформированных на базе изучаемой генеральной совокупности

Таким образом, что же имеем на самом деле? Работаем с одной выборкой, которая характеризуется средним $\bar{x}$ и стандартным отклонением S_n. Их то и можно использовать для прогнозирования генеральной средней $\tilde{x}$ путем расчета случайной ошибки $S_{\bar{x}}$ с последующим определением доверительного интервала Δx (рис.1.2). При этом важно отметить и надежность наших поисков, для чего надлежит указать также доверительную вероятность сделанных заключений.

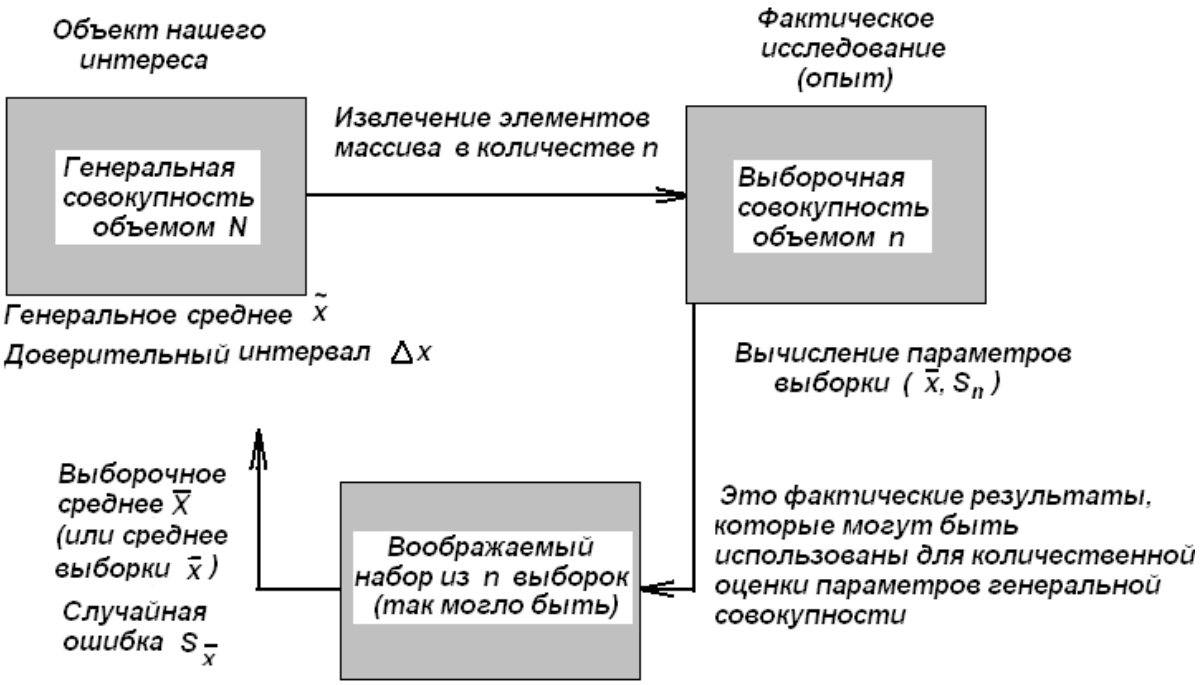


Рис.1.2. Схема выборочного исследования

1.7. О доверительной вероятности и доверительном интервале. Понятие о предельной ошибке

Напомним об упомянутых здесь таких понятиях, как доверительная вероятность и доверительный интервал.

Вероятность можно рассматривать как средство для работы в условиях риска и неопределенности. Она показывает возможность (или шанс) наступления в будущем каждого из различных потенциальных событий, рассчитанную на основании информации о некоторой ситуации.

Вероятность – это понятие, в некотором смысле обратное статистике. Если *статистика* помогает переходить от *наблюдений* к *обобщениям* относительно рассматриваемой ситуации, то *вероятность* имеет обратную направленность. А именно: исходя из характеристики ситуации, можно выяснить, *какие данные* мы, скорее всего, *получим* и *какова возможность получения этих данных*.

Основной закон теории вероятности – закон больших чисел – утверждает, что при достаточно большом числе измерений (наблюдений) N частота появления $f_N(A)$ некоторого события A как угодно мало отличается от вероятности этого события $P(A)$:

$$| P(A) - f_N(A) | < \varepsilon,$$

где ε – сколь угодно малое положительное число, отличное от нуля ($\varepsilon > 0$).

Частота события – это активность (интенсивность) проявления того, что имеет место быть (наступление реального события).

Вероятность события – это предположение (прогноз) о возможном наступлении этого события.

В случае событий массового характера вероятность может рассматриваться как мера объективной возможности наступления события. Около числа $P(A)$ группируются относительные частоты события A.

Пусть среднее арифметическое для случайной выборки равно $\bar{x}$, а среднее арифметическое для генеральной совокупности – $\tilde{x}$.

Примем, что P означает вероятность того, что результат измерения среднего для выборки (мы его знаем, т.к. можем сосчитать) отличается от среднего генеральной совокупности (этого мы не знаем, но хотим знать) на величину, не большую чем Δx.

Это условие можно записать так:

$$p(-\Delta x < \tilde{x} - \bar{x} < \Delta x) = P.$$

Здесь вероятность P носит название *доверительной вероятности* (или коэффициента надежности). Интервал значений от $\bar{x} - \Delta x$ до $\bar{x} + \Delta x$ (или $\pm\Delta x$) называется *доверительным интервалом*.

Доверительный интервал – это интервал, внутри которого с заданной степенью достоверности (надежности) находится значение искомого параметра (в данном случае среднее генеральной совокупности $\tilde{x}$).

Пояснение. Вычислив доверительный интервал, мы может утверждать, что с указанной надежностью (вероятностью) генеральное среднее $\tilde{x}$ отличается от среднего выборки $\bar{x}$ на величину, не превышающую этот интервал. Иначе говоря, $\tilde{x}$ лежит где-то внутри доверительного интервала. Но это утверждение делается с оговоркой – это верно, но с такой-то вероятностью.

Длительный опыт применения статистических расчетов показал, что наиболее приемлемой величиной доверительной вероятности является 95 %. Однако используют и другие показатели: 90, 99 и даже 99,9 %. Уровень 95 % представляет собой определенный компромисс между попыткой, с одной стороны, получить по возможности более высокий уровень надежности и, с другой, желанием иметь относительно небольшой интервал. Платой за более высокую доверительную вероятность является более широкий и, значит, менее полезный доверительный интервал.

Другим показателем меры наших требований к статистическому исследованию является *уровень значимости* (или *уровень риска*) α:

$$\alpha = 1 - P \text{ или } P = 1 - \alpha.$$

Часто используется $\alpha = 0{,}05$; это значение, называемое еще 5 %-*ным уровнем риска*, соответствует вероятности верного утверждения, равного $P = 1 - \alpha = 0{,}95$ или 95 %.

Обычно используют следующие фразы для описания результатов (табл. 1.1):

Т а б л и ц а 1.1

Интерпретация количественных значений уровня значимости α

Описание	Интерпретация
Незначимый ($\alpha > 0{,}05$)	Незначимый на обычном уровне 5%
Значимый ($\alpha < 0{,}05$)	Значимый на обычном уровне 5%, но незначимый на уровне 1%
Высоко значимый ($\alpha < 0{,}01$)	Является значимым на уровне 1%, но незначимым на уровне 0,1%
Очень высоко значимый ($\alpha < 0{,}001$)	Значимый на уровне 0,1%

Полезно запомнить: параметр генеральной совокупности $\tilde{x}$ находится между значением оценки (средним арифметическим $\bar{x}$) в интервале:

± стандартная ошибка при доверительной вероятности 68 %;

± 2·(стандартная ошибка) при доверительной вероятности 95,4 %;

± 3·(стандартная ошибка) при доверительной вероятности 99,7 %.

Для корректного построения доверительного интервала необходимо выполнение двух условий*:*

– выборка должна быть *случайной;*

– распределение должно быть *нормальным.*

В заключение выскажем одно соображение. Определяя доверительный интервал, мы тем самым указываем на ту погрешность, с которой вычисляем истинное значение совокупности. Однако ценность этой информации практически теряется, если при этом не указывать величину достоверности, с которой найден искомый результат. Таким образом, для характеристики величины ошибки нужно задать два числа: *а*) величину самой погрешности, т.е. доверительного интервала, и *б*) величину доверительной вероятности.

1.8. Критерий Стьюдента

В статистике принято пользоваться понятием *степень свободы*. Под этим понимают число независимых элементов информации, которые взяты для вычисления стандартной ошибки. Для одной выборки число степеней свободы равно $n - 1$ (число на единицу меньше количества наблюдений или элементов массива). Или иначе: это разность между числом измерений (наблюдений) и числом коэффициентов (констант), которые уже вычислены по результатам этих измерений. В данном случае по результатам измерения случайной выборки мы рассчитали среднее арифметическое (это наша константа).

При обработке данных, количество которых ограничено, принято при использовании стандартной ошибки вводить специальный корректирующий показатель – *критерий* или *коэффициент Стьюдента* (*t-критерий*).

Применение *t*-критерия основано на знании особенностей распределения при ограниченном числе наблюдений (малой выборке). В распределении Стьюдента максимум частоты совпадает с максимумом частоты нормального распределения, но высота и ширина кривых зависят от числа элементов, входящих в выборочную совокупность. Чем меньше число измерений $\boldsymbol{n}$, тем более пологий ход имеет кривая распределения. При $\boldsymbol{n} \geq 20$ распределение Стьюдента переходит в нормальное распределение (рис.1.3).

Величина доверительного интервала определяется по формуле:

$$\Delta x = \frac{t_{\alpha;n} \cdot S_n}{\sqrt{n}} \quad \text{или} \quad \boldsymbol{\Delta x = t_{\alpha,n} S_{\bar{x}}}$$

В статистике полученную величину Δx принято называть также *предельной ошибкой выборки*. Тогда величина среднего генеральной совокупности будет определяться следующим выражением

$$\tilde{x} = \bar{x} \pm \Delta x \text{ или } \tilde{x} = \bar{x} \pm t_{\alpha,n} S_{\bar{x}}.$$

Здесь множитель t (критерий Стьюдента) в статистике называется также *коэффициентом доверия.* Он определяется в зависимости от того, с какой доверительной вероятностью надо гарантировать результаты выборочного обследования.

Видно, что $\Delta x = \pm S_{\bar{x}}$ при t =1, т.е. для P=0,68.

Фактически доверительный интервал (предельная ошибка) Δx – это та же случайная ошибка $S_{\bar{x}}$, но только кратно (на величину t) отличающаяся от нее. Следовательно, критерий Стьюдента рассматривается как коэффициент кратности стандартной ошибки.

Рис.1.3. Кривые нормального распределения (1) и распределения Стьюдента (2)

Значения t-критерия выбираются по специальным статистическим таблицам в зависимости от *а*) доверительной вероятности P (или уровня значимости α) и *б*) числа измерений n (см. *Приложение 1*).

Для *повторного* отбора используется упомянутое выше выражение

$$\Delta x = t_{\alpha,n} S_n / \sqrt{n} \text{ или } \Delta x = t_{\alpha,n} S_{\bar{x}}.$$

Для случая *бесповторного* отбора:

$$\Delta x = t \cdot S_n \sqrt{\frac{(1 - n/N)}{n}}.$$

При малом объеме единиц совокупности, взятых в выборку (обычно <5%), множитель (1- n/N) близок к 1. Поэтому в упрощенном варианте вновь имеем $\Delta x = t_{\alpha,n} S_n / \sqrt{n}$ (т.е. как для повторного отбора).

1.9. Необходимое число измерений (оптимальный объем выборки)

Показано, что величина доверительного интервала (предельной ошибки) зависит от *а*) объема выборки и *б*) степени вариации (изменчивости) признака, выраженной через дисперсию.

Уменьшение ошибки, и, следовательно, повышение точности оценки всегда связано с увеличением объема выборки. Поэтому уже на стадии организации выборочного наблюдения приходится решать вопрос о том, каков должен быть объем выборочной совокупности, чтобы была обеспечена требуемая точность результатов наблюдения.

При формировании объема выборки можно придерживаться общего подхода, полагая, что она должна составлять 5-10% (реже 15-25%) от объёма генеральной совокупности.

Однако такой отбор не позволяет судить о степени достоверности будущих результатов (доверительной вероятности). Кроме того, бывают ситуации, когда на счету, что называется, каждая единица массива и возможный "перебор" приведет к незапланированным чрезмерным расходам.

Известны различные рекомендации для определения необходимого числа элементов выборки, чтобы получить результат исследования с заданной вероятностью.

Так, используется следующая формула для расчета численности выборки n_x:

для *бесповторного* отбора $n_x = \frac{N \cdot t^2 \cdot \sigma^2}{N \cdot \Delta x^2 + t^2 \cdot \sigma^2}$;

для *повторного* отбора $n_x = \frac{t^2 \cdot \sigma^2}{\Delta x^2}$,

где t –критерий Стьюдента, σ^2 – дисперсия генеральной совокупности, N – размер генеральной совокупности; Δx – доверительный интервал (предельная ошибка).

Особенность представленных формул в том, что в первом случае можно вести расчет, отталкиваясь от известного нам объема самой генеральной совокупности N. Вторая формула позволяет получить результат, формально игнорируя её количественный размер.

При планировании выборочного исследования предполагается заранее, что известны следующие данные:

– величина допустимой ошибки выборки Δx (доверительного интервала);

– вероятность выводов по результатам наблюдения (величина t-критерия при заданной доверительной вероятности P или уровне значимости α).

Величина σ^2, характеризующая дисперсию признака в генеральной совокупности, чаще всего бывает неизвестна. Поэтому используют следующие приближенные способы оценки генеральной дисперсии.

1. Можно провести пробное исследование (обычно небольшого объема), на базе которого определяется величина дисперсии этой выборки, используемой в качестве оценки генеральной дисперсии:

$$\sigma^2 = \frac{\sum (x_i - \bar{x}_{проб})^2}{n_{проб} - 1},$$

где $\bar{x}_{проб}$ – среднее арифметическое по результатам пробного исследования; $n_{проб}$ – число единиц, попавших в пробное исследование.

По данным нескольких таких маломасштабных экспериментов выбирается наибольшее значение дисперсии, которое и будет использовано при проведении полного исследования.

2. Можно использовать данные прошлых выборочных наблюдений, проводившихся в аналогичных целях, т.е. дисперсия, полученная по их результатам, применяется в качестве оценки генеральной дисперсии.

3. Если распределение признака в генеральной совокупности может быть отнесена к нормальному закону распределения, то размах вариации примерно равен 6σ (крайние значения отстоят в ту и другую сторону от средней на расстоянии 3σ для P=99,7%), т.е. $R=6\sigma$, откуда $\sigma \cong 1/6R$, где $R = x_{max} - x_{min}$.

При выполнении статистических исследований с коммерческими целями можно практически с достаточной точностью указать максимально и минимально возможные значения исследуемого параметра (признака) в анализируемой совокупности.

Рассмотрим примеры.

Пример 1. *Руководство Сбербанка решило выяснить, каков средний уровень оплаты труда операционистов (банковских клерков, непосредственно работающих с клиентами) во всех отделениях Банка Уральского Федерального Округа. Аналитический отдел приступил к этой работе. При этом для большей информированности было решено провести такой расчет, принимая во внимание а) степень достоверности (доверительную вероятность* ***P****) и б) точность оценки (доверительный интервал* ***Δx****).*

Предварительным обследованием было установлено, что различие между наивысшим и наименьшим уровнем оплаты труда операциониста в регионе составляет 6000 руб.

Корме того, руководство интересовало также, каким образом изменится объем выборки при различной величине предельной ошибки, а именно: при ***Δx****, равной соответственно 100, 200 и 500 руб.*

> ***Пояснение.*** Напомним, что для нормального распределения в промежуток $\Delta x = \bar{x} \pm 3\sigma$ включается 99,7% всех вариантов значений параметра. Применительно к нашей задаче это означает, что 6000 руб. примерно равно шести стандартным отклонениям (6000 = 6σ). Для $\Delta x = \bar{x} \pm 2\sigma$ доверительная вероятность составит 95,4% . Наконец, для $\Delta x = \bar{x} \pm \sigma$ – это 68,3% .

Все необходимые расчеты исполним посредством программы Excel. Будем действовать в следующей последовательности.

1. Запустим Excel и откроем рабочий лист, которому можно придумать подходящее наименование. Не мудрствуя лукаво, отметим наш документ скромным именем "Выборка".
2. Теперь введем исходные данные. Для этого в ячейку A1 поместим надпись "Размах ***R***", в A2 – "***t***-критерий", в A3 – "Стандартное отклонение σ", в A4 – "Доверительный интервал Δx" и в A5 – "Объем выборки ***n***".
3. В колонке B зарезервируем ячейки, соответствующие нашим параметрам, указанным в колонке A. Укажем в ячейке B1 значение размаха, равного 6000, а в ячейке B3 стандартное отклонение σ, равное 1000.

Затем запишем формулу, по которой будем считать объем выборки (для случая повторного отбора), т.е.

$$n_x = \frac{t^2 \cdot \sigma^2}{\Delta x^2} .$$

Для этого выделим ячейку B5 и в поле ввода формул поместим последовательно знак равенства и саму формулу, указывая необходимые операторы (знаки математических действий) и ссылки на соответствующие ячейки. Затем укажем данные для случая, когда t =1 и Δx =100 (начнем расчет с этой комбинации).
После этого активизируем ячейку B5, в которой и появится рассчитанное значение объема выборки. Оно равно 100 (рис.1.4).

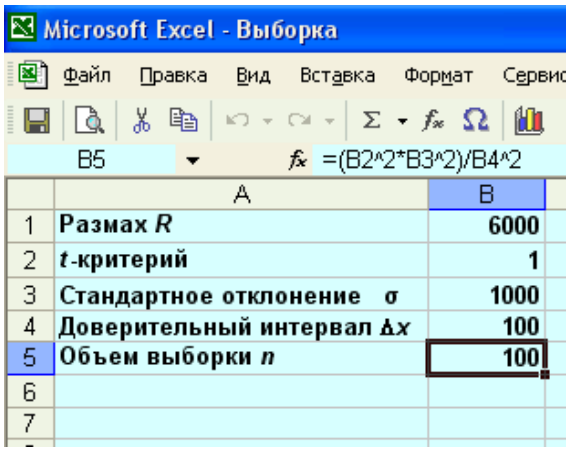
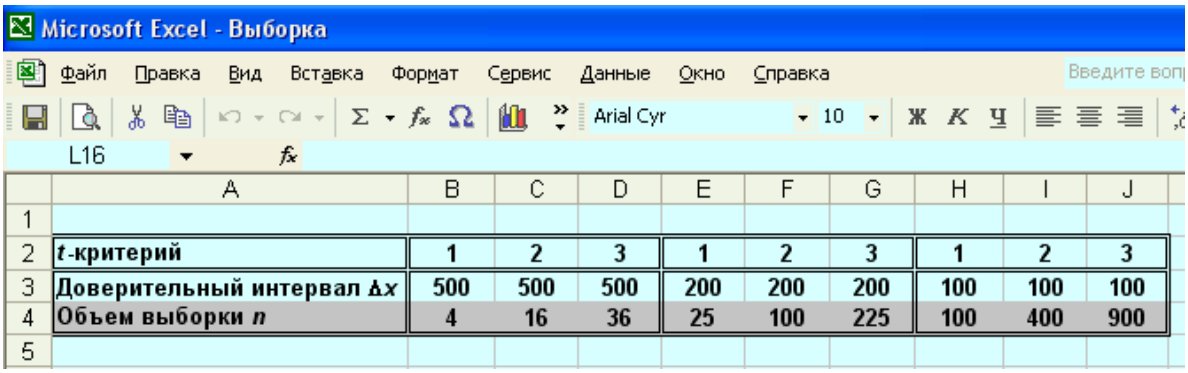


Рис.1.4. Лист Excel с исходными данными и результатом расчета объема выборки

4. Теперь последовательно будем вводить в ячейки B2 и B4 наши данные, перебирая заданные значения t и Δx, а в ячейке B5 станем считы-

вать новые значения ***n***. Для удобства организуем таблицу, в которой поместим полученные результаты (число операционистов, для которых затем нужно будет рассчитать средний уровень жалования). Для этого перейдем на другой лист и итоговый вид таблицы можно видеть на рис.1.5.

Собственно говоря, на этом наши вычисления закончены. Для более удобного анализа полученных результатов итоговую таблицу можно представить несколько по-иному – указать не сами значения ***t***-критерия, а величины соответствующей доверительной вероятности ***P*** (табл.1.2).



	A	B	C	D	E	F	G	H	I	J
1										
2	***t*-критерий**	**1**	**2**	**3**	**1**	**2**	**3**	**1**	**2**	**3**
3	**Доверительный интервал Δ*x***	**500**	**500**	**500**	**200**	**200**	**200**	**100**	**100**	**100**
4	**Объем выборки *n***	**4**	**16**	**36**	**25**	**100**	**225**	**100**	**400**	**900**
5										

Рис.1.5. Лист Excel с результатами расчета

Т а б л и ц а 1. 2

Объем выборки в зависимости от уровня достоверности *P* и точности оценки Δ*x*

Доверительная вероятность *P*	**Предельная ошибка (доверительный интервал) Δ*X*, руб.**		
	100	**200**	**500**
0,683	**100**	**25**	**4**
0,954	**400**	**100**	**16**
0,997	**900**	**225**	**36**

Резюме. Если ориентироваться на наиболее принятую величину доверительной вероятности, равную 95,4%, то в зависимости от заявленной точности измерения (100, 200 или 500 руб.) нужно будет сформировать выборку объемом соответственно в 400, 100 и 16 человек. А это уже повод для раз-

мышления высокого начальства Сбербанка – какой результат по точности и достоверности требуется. Если рискнуть на высокую точность, равную 100 руб., то нужна достаточно приличная по размеру выборка – 400 человек. Весьма грубая оценка (±500 руб.) будет связана с минимальными затратами на такое обследование (объем выборки будет составлять всего-то 16 единиц). Наконец, компромиссный вариант – при погрешности в 200 руб. потребуется обсчитать 100 человек. Наверное, это наиболее приемлемый вариант.

Пример 2. *Фирма, занимающаяся производством гвоздей и шурупов, заказала у своего поставщика, метизно-металлургического завода, 120 мотков стальной проволоки нужных диаметров. В соответствии с согласованными техническими требованиями вес каждого мотка должен составлять не менее 60 кг, при этом допускается отклонение от этой величины не более чем на 5% (т.е. погрешность ± 3 кг). Отделом снабжения фирмы решено было провести контрольные измерения весовых показателей закупленной продукции, чтобы убедиться в добросовестности и надежности своего торгового партнера.*

Необходимо рассчитать, сколько нужно взвесить мотков из этой партии, чтобы быть уверенным в соблюдении указанных условий с вероятностью 90 и 95%.[*] *Установлено, что дисперсия* σ^2 *составляет 31,4.*

При решении этой задачи придется формировать выборку без возврата (нет нужды вновь перевешивать какой-то моток, если он вновь случайно оказался подлежащим извлечению). Поэтому воспользуемся формулой для бесповторного отбора:

$$n_x = \frac{N \cdot t^2 \cdot \sigma^2}{N \cdot \Delta x^2 + t^2 \cdot \sigma^2}.$$

Прежде всего, определим, какие табличные значения ***t***-критерия будут соответствовать указанным вероятностям. Для этого воспользуемся эталонной таблицей (*Приложение 1*) и найдем, что для ***P***=0,90 (или α=0,1) и ***N***=120 табличное значение ***t***-критерия составит 1,658, а для ***P***=0,95 (или α=0,05) – соответственно 1,980.

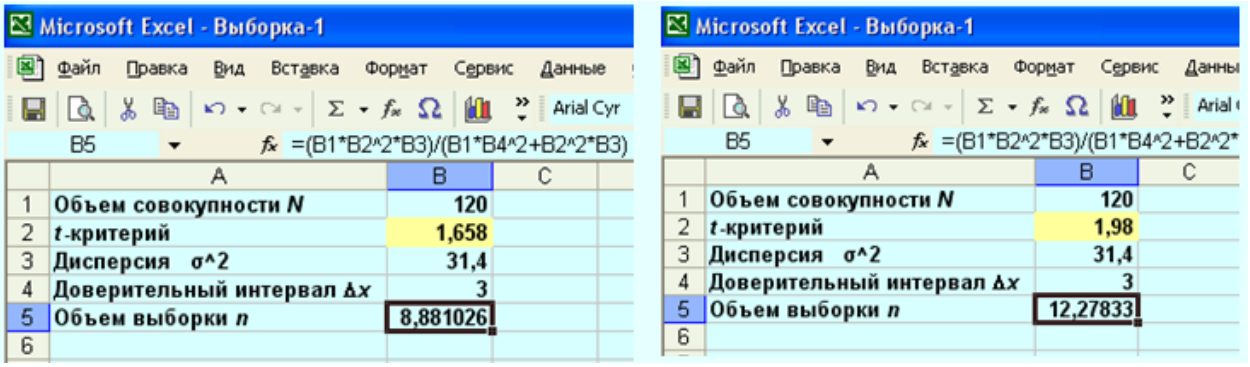
[*] В статистических расчетах довольно часто для удобства используются округленные значения вероятности, что, естественно, будет отвечать определенным показателям *t*-критерия. Так, для вероятности, равной 90, 95 или 99%, и при $n = \infty$ значения *t*-критерия составят 1,645, 1,960 и 2,576 соответственно (см. *Приложение 1*).

Теперь вновь запустим Excel. Откроем тот же файл "Выборка", только активизируем новое окно (Лист 3). Поскольку пользование программой оказывается аналогичным рассмотренному выше примеру, ограничимся лишь краткими пояснениями (рис.1.6).

1. В ячейках A1:A5 построчно запишем нужные наименования: "Объем совокупности ***N***", "***t***-критерий", "Дисперсия σ^2", "Доверительный интервал Δx" и "Объем выборки ***n***".
2. Затем в ячейках B1:B4 укажем соответствующие числовые данные, при этом расчет начнем с варианта, когда ***P***=0,90, чему соответствует значение ***t***=1,658.
3. Активизируем ячейку B5 и в поле ввода запишем формулу для расчета ***n***. Сместим курсор в эту ячейку (появится белый крестик) и щелкнем левой клавишей – в ячейке B5 фиксируется рассчитанное значение ***n***, равное 8,881026 (рис.1.6*а*). Затем в ячейку B2 запишем число 1,98 и в ячейке прочитаем новый показатель – 12,27833 (рис.1.6*б*).

На этом расчет закончен. С учетом округления до целых чисел получим 9 и 12.

Таким образом, чтобы выполнить заданные условия, нужно будет проверить вес соответственно 9 и 12 мотков.



а б

Рис.1.6. Результаты расчета количества мотков для доверительной вероятности 90 (*а*) и 95% (*б*)

1.10. Случайная выборка

Итак, мы умеем определять необходимое число измерений, чтобы в дальнейшем можно было бы на основании этой информации получать вполне достоверные сведения обо всей совокупности в целом. Теперь нужно научиться отбирать сами элементы массива для формирования выборки заданного объема. Иначе говоря, мы знаем, *сколько* извлечь, теперь нужно научиться, *как* это сделать.

Чтобы избежать какой-либо тенденциозности и предвзятости при отборе, формирование выборки должно осуществляться случайным образом.

Случайная выборка состоит в том, что *а*) каждый элемент генеральной совокупности имеет *одинаковую вероятность быть отобранным* и *б*) элементы *отбираются независимо* друг от друга.

Независимость отбора обеспечивает сбор максимально возможного объема независимой информации и выше, следовательно, будет вероятность репрезентативности.

1.10.1. Таблица случайных чисел

Один из способов извлечения случайной выборки – *применение таблицы случайных чисел*.

Таблица случайных чисел представляет собой организованную в виде таблицы последовательность цифр, в которой каждая из цифр от 0 до 9 встречается независимо друг от друга с вероятностью 1/10.

Существуют разные по конструкции таблицы случайных чисел. Они могут представлять наборы из 2-х, 3-х, 4-х или 5-ти случайных цифр, расположенных в произвольном порядке.

Табл.1.3 дает представление о конструкции такой таблицы, построенной из комбинаций 4-х случайных цифр.

Для получения случайной выборки путем отбора без возврата принято пользоваться следующим правилом.

1. Предварительно нужно составить основу выборки (список) таким образом, чтобы все элементы генеральной совокупности были пронумерованы числами от 1 до N.

2. Выбрать точку начала считывания случайных чисел из таблицы. Это необходимо сделать случайным образом, например, подбросив монету.

3. Начав с выбранной точки, последовательно считывать цифры слева направо с переходом на следующую строку.

4. Объединить эти цифры в группы, размер которых равен количеству цифр в числе ***N***. Так, если ***N*** трехзначное число, то нужно считывать по три случайные цифры раз за разом.

5. И таким образом поступать, пока не получится выборка из ***n*** единиц, придерживаясь следующих рекомендаций:

– если получилось случайное число в диапазоне от 1 до ***N*** и элемент с таким номером еще не извлекался, то его нужно включить в выборку;

– если полученное случайное число равно 0 или больше ***N***, то его нужно проигнорировать, поскольку для него в основе выборки нет соответствующего элемента генеральной совокупности;

– если окажется, что элемент с таким номером уже извлекался, то его следует пропустить, поскольку осуществляется выборка без возврата.

Проиллюстрируем пользование таблицей на конкретном примере.

Воспользуемся рассмотренным ранее случаем с формированием выборки из мотков проволоки. Возьмем вариант с объемом ***n***=12.

Итак, размер генеральной совокупности ***N***=120. Составим список всех мотков, которым присвоены номера от 1 до 120. Допустим, случайным образом было решено начать отсчет с шестой комбинации цифр, т.е. с числа 3912 (см. табл.1.3. строка 1, столбец 6). Поскольку число ***N***=120 состоит из трех цифр, объединим последовательность случайных чисел в группы, состоящие также из трех цифр, следующим образом: 391 209 387 460 086 944 203 522 093… Отбросим первые четыре комбинации из этого ряда, т. к. они более 120. Первым попавшим в выборку будет число 086. Затем опять пропускаем несколько чисел, пока не встретится комбинация 093. И так далее…

Процесс продолжается, пока не будет отобрано ***n***=12 элементов. Как, видно, ими окажутся мотки с номерами 086, 093, 095, 090, 001, 043, 038, 104, 012, 050, 043 и 105. Вот они то и составят нашу случайную выборку.

Т а б л и ц а 1. 3

Таблица случайных чисел

5489	5583	3156	0835	1988	3912	0938	7460	0869	4420
3522	0935	7877	5665	7020	9555	7375	7124	7878	5544
7555	7579	2550	2487	9477	0864	2349	1012	8250	2633
5759	3554	5080	9074	7001	6249	3224	6368	9102	2672
6303	6895	3371	3196	7231	2918	7380	0438	7547	2644
7351	5634	5323	2623	7803	8374	2191	0464	0696	9529
7068	7803	8832	5119	6350	0120	5026	3684	5657	0304
3613	1428	1796	8447	0503	5654	3254	7336	9536	1944
5143	4534	2105	0368	7890	2473	4240	8652	9435	1422
9815	5144	7649	8638	6137	8070	5345	4865	2456	5708
5780	1277	6816	1013	2867	9938	3930	3203	5696	1769
1187	0951	5991	5245	5700	5564	7352	0891	6249	6568
4184	2179	4554	9083	2254	2435	2965	5154	1209	7069
2916	2972	9885	0275	0144	8034	8122	3213	7666	0230
5524	1341	9860	6565	6981	9842	0171	2284	2707	3008
0146	5291	2354	5694	0377	5336	6460	9585	3415	2358
4920	2826	5238	5402	7937	1993	4332	2327	6875	5230
7978	1947	6380	3425	7267	7285	1130	7722	0164	8573
7453	0653	3645	7497	5969	8682	4191	2976	0361	9334
1473	6938	4899	5348	1641	3652	0852	5296	4538	4456
8162	8797	8000	4707	1880	9660	8446	1883	9768	0881
5645	4219	0807	3301	4279	4168	4305	9937	3120	5547
2042	1192	1175	8851	6432	4635	5757	6656	1660	5389
5470	7702	6958	9080	5925	8519	0127	9233	2452	7341
4045	1730	6005	1704	0345	3275	4738	4862	2556	8333
5880	1257	6163	4439	7276	6353	6912	0731	9033	5294
9083	4260	5277	4998	4298	5204	3965	4028	8936	5148
1762	8713	1189	1090	8989	7273	3213	1935	9321	4820
2023	2589	1740	0424	8924	0005	1969	1636	7237	1227
7965	3855	4765	0703	1678	0841	7543	0308	9732	1289
7690	0480	8098	9629	4819	7219	7241	5128	3853	1921
9292	0426	9573	4903	5916	6576	8368	3270	6641	0033
0867	1656	7016	4220	2533	6345	8227	1904	5138	2537
0505	2127	8255	5276	2233	3956	4118	8199	6380	6340
6295	9795	1112	5761	2575	6837	3336	9322	7403	8345
6323	2615	3410	3365	1117	2417	3176	2434	5240	5455
8672	8536	2966	5773	5412	8114	0930	4697	6919	4569
1422	5507	7596	0670	3013	1351	3886	3268	9469	2584
2653	1472	5113	5735	1469	9545	9331	5303	9914	6394
0438	4376	3328	8649	8327	0110	4549	7955	5275	2890
2851	2157	0047	7085	1129	0460	6821	8323	2572	8962
7962	2753	3077	8718	7418	8004	1425	3706	8822	1494
3837	4098	0220	1217	4732	0150	1637	1097	1040	7372
8542	4126	9274	2251	0607	4301	8730	7690	6235	3477
0139	0765	8039	9484	2577	7859	1976	0623	1418	6685
6687	1943	4307	0579	8171	8224	8641	7034	3595	3875
6242	5582	5872	3197	4919	2792	5991	4058	9769	1918
6859	9606	0522	4993	0345	8958	1289	8825	6941	7685
6590	1932	6043	3623	1973	4112	1795	8465	2110	8045
3482	0478	0221	6738	7323	5643	4767	0106	2272	9862

1.10.2. Метод механического отбора

Помимо использования таблицы случайных чисел другим распространенным приемом в практике выборочного наблюдения является *механический (периодический) отбор*. Иногда формируемую этим методом выборку называют *систематической*. Для её получения из генеральной совокупности извлекаются *такие элементы*, которые находятся в массиве на *равном расстоянии* друг от друга.

Допустим, имеется полный список единиц совокупности и эти единицы располагаются в порядке, являющемся случайным по отношению к подлежащим изучению признакам (например, список сотрудников фирмы по алфавиту). В зависимости от объема выборки из списка для обследования выбирается каждая четвертая единица или каждая десятая. При проведении механической выборки генеральная совокупность фактически разбивается *на равные по численности группы (интервалы)* и из каждой такой группы *отбирается одна единица*.

В том случае, когда к механическому отбору прибегают с целью повышения репрезентативности, списки единиц генеральной совокупности составляют в форме *ранжированного ряда* (по возрастанию или убыванию какого-то признака). Так, при изучении бюджета служащих фирмы используется механический отбор из списков, составленных в порядке убывания величины средней месячной зарплаты.

Механический отбор полезен и тогда, когда *невозможно заранее составить список* элементов массива. Например, выборка берется из совокупности

а) *постепенно* формирующейся во времени или

б) практически *бесконечной* совокупности.

Так, при исследовании процесса покупки можно наблюдать каждого десятого покупателя; при контроле качества продукции проверять, например, каждую пятую сходящую со станка деталь и т.д.

При проведении механической выборки нужно выполнить следующие процедуры:

1. Установить *шаг отсчета (размер интервала)* h, т.е. выбрать расстояние между отбираемыми единицами.

Шаг устанавливают в зависимости от предполагаемого процента отбора. Его размер равен *обратной величине доли выборки*. Так, при 2%-ной выборке отбирается каждая 50-я единица (1:0,02), при 5%-ной выборке (1:0,05) – каждая 20-я единица и т.д.

Допустим, из генеральной совокупности объёмом 1000 единиц обследованию подлежат 100 элементов (т.е. 10%). Это значит, что из каждых 10 единиц обследование пройдет только одна единица. Следовательно, шаг отсчета равен 10.

Это означает, что шаг можно определить как отношение ***h=N/n.***

2. Выбрать *начало отсчета*, т.е. номер той единицы, которая должна быть обследована первой.

Выбор начала отсчета связан со способом расположения единиц генеральной совокупности в списках. В случае *неупорядоченного* расположения единиц из совокупности единиц первого интервала путем случайного отбора выбирают начальную единицу. Можно подбросить монетку, провести жеребьевку – из шляпы вынуть бумажку с соответствующим номером (вспомним впечатляющий эпизод из кинофильма "Гараж"). Предположим, что для случая отбора 100 элементов из массива в 1000 единиц в результате жеребьевки номер начальной единицы составил 4. Тогда в выборку попадут элементы массива, стоящие в списке под номерами 4, 14, 24, 34,…, 984, 994.

Если элементы в списке были *ранжированы*, то за начало отсчета принимают единицу, лежащую в середине первого интервала. В данном примере из первых десяти единиц нужно выбрать пятую или шестую единицу, Тогда в выборку попадают единицы с порядковыми номерами 5, 15, 25, 35,…, 985 и 995 (или же 6, 16, 26, 36, …, 986 и 996).

1.11. Компьютерное формирование выборочной совокупности

Выше нами был рассмотрен прием извлечения выборки "вручную". Познакомимся теперь с компьютерным методом выполнения этой процедуры на основе применения приложения Excel. Воспользуемся программой ***Анализ данных***, в которую вложен инструмент ***Выборка***. Реализуются две методики – с *повторным* отбором (с возвращением) и с *бесповторным* отбором (без возвращения).

1.11.1. Повторный отбор

Для удобства рассмотрим случай с уже знакомыми нам мотками проволоки. Напомним, что генеральная совокупность, подлежащая изучению, насчитывает 120 элементов (мотков проволоки). В соответствии с полученными расчетами для доверительной вероятности P=0,95 и доверительного интервала Δx=±3 объем случайной выборки n должен составить 12 единиц.

Итак, случайным образом нам надлежит сформировать выборку именно такого размера.

Полагаем, что все мотки нами поименованы – им присвоены номера от 1 до 120.

1. Запускаем Excel и в имеющемся листе (Лист 4) укажем вначале заголовок "Выборка с возвратом". Текст довольно длинный, он захватывает несколько ячеек – A1, B1 и C1. Чтобы заголовок удобно располагался, лучше эти ячейки объединить. Для этого в главном меню выберем опции ***Формат/Ячейки…*** В появившемся окне диалога ***Формат ячеек*** активизируем вкладку ***Выравнивание*** и в списке ***Отображение*** отметим флажком ***Объединение ячеек***. Схожим образом поступим с заголовками "Номер мотка" и "Выборка случайная (***n***=12)". Они займут соответственно ячейки A3:B3 и C3:E3.

2. Поместим затем номера мотков в ячейки A4:A123 (рис.1.7). Вводить последовательно все числа в свои ячейки – занятие довольно занудное. Однако есть весьма элегантный способ. Чтобы быстро ввести числа от 1 до 120, поступим следующим образом. В ячейки A4 и A5 введем цифры 1 и 2. Затем выделим эти ячейки и протянем маркер заполнения (черный квадратик в правом нижнем угле) вниз столбца A, следя за счетчиком заполняемых ячеек (он появится справа от маркера). Остановимся, когда счетчик укажет число 120.

3. В меню ***Сервис*** выберем ***Анализ данных*** и в открывшемся окне ***Инструменты анализа*** выделим опцию ***Выборка***, после чего – ***ОК***.

4. В появившемся окне ***Выборка*** укажем диапазон входящих данных. Для этого в текстовом поле ***Входной интервал*** отметим диапазон ячеек рабочего листа A3:A123 (вместе с заголовком). Поэтому установим флажок ***Метки.***

5. Укажем метод отбора, а именно: ***Случайный***. Отметим также нужный нам объем выборочной совокупности – ***Число выборок*** (оно равно 12**),** а также ячейку (C4), в которую будет помещен полученный результат – это ***Выходной интервал***. После чего ***ОК*** (рис. 1.7).

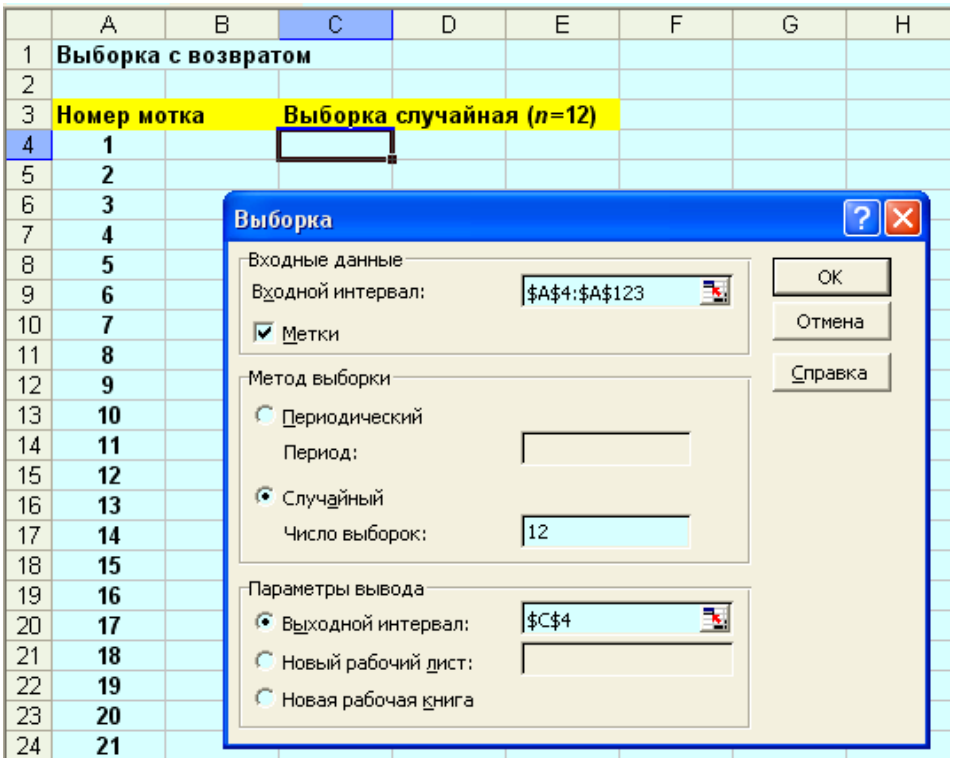


Рис.1.7. Исходные данные и диалоговое окно *Выборка*

Полученный результат показан на рис. 1.8. Как видно, в случайную выборку попадают мотки проволоки под номерами (в порядке возрастания) 6, 10, 14, 28, 31, 48, 66, 68, 91, 93, 96, и 118.

	A	B	C	D	E
1	Выборка с возвратом				
2					
3	Номер мотка		Выборка случайная (n=12)		
4	1		118		
5	2		93		
6	3		96		
7	4		31		
8	5		14		
9	6		10		
10	7		68		
11	8		91		
12	9		48		
13	10		6		
14	11		28		
15	12		66		
16	13				
17	14				

Рис.1.8. Результаты формирования случайной выборки

Пояснение. Не исключено, что могут случаться повторы (ибо рассматривался способ с возвращением). Как быть? Если это принципиально важно, то можно повторять отбор, пока в выборке не окажутся только неповторяемые номера.

Теперь попробуем выполнить извлечение из этого же массива в режиме механического (периодического) отбора. Установим шаг ***h***, пусть он составит 10, тогда из совокупности в 120 единиц нужно будет отобрать те же 12 мотков. В диапазоне ячеек C18:F18 запишем "Выборка периодическая (шаг ***h***=10)". Далее будем действовать знакомым образом.

6. В диалоговом окне ***Выборка*** воспользуемся опцией ***Периодический*** и в текстовом поле ***Период*** укажем шаг, с которым должны извлекаться значения из исходного массива данных. Затем отметим ячейку (C20), в которую будет помещен полученный результат – это ***Выходной интервал***. После чего ***ОК*** (рис. 1.9).

В данном случае в выборку попадают мотки проволоки под номерами 11, 21, 31, …, 101 и 111. (рис.1.10). Начало отсчета Excel организует случайным образом.

Таким образом, мы сумели сформировать случайную выборку, используя разные способы компьютерного извлечения из генеральной совокупности.

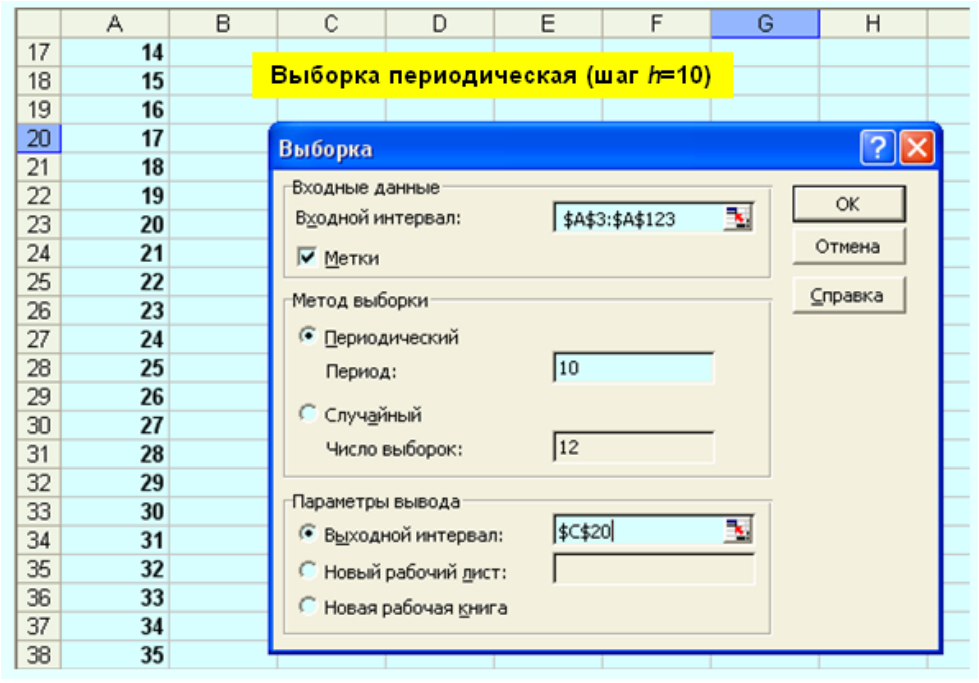


Рис.1.9. Формирование выборки способом механического отбора

	A	B	C	D	E	F
16	13					
17	14					
18	15					
19	16					
20	17		11			
21	18		21			
22	19		31			
23	20		41			
24	21		51			
25	22		61			
26	23		71			
27	24		81			
28	25		91			
29	26		101			
30	27		111			
31	28					

Рис.1.10. Результаты формирования выборки механическим отбором

В практической статистике обычно принято отдавать предпочтение случайной выборке. Дело в том, что может обернуться серьезной неудачей применение систематической выборки в том случае, если в основе выборки существует определенный повторяемый фрагмент, который по размеру соответствует шагу отбора. Скажем, при сборке на конвейере каждому 20-му холодильнику уделяется особое внимание (не будем предполагать, почему именно). Тогда по воле случая может оказаться, что механическим отбором в выборку извлекаются именно эти 20-е. Понятно, что результаты окажутся вполне бессмысленными, поскольку о репрезентативности такого выборочного массива говорить не приходится.

1.11.2. Бесповторный отбор

Во многих случаях возникает ситуация, когда нужно получить такую выборку, чтобы каждое значение, извлеченное из генеральной совокупности, встречалось не более одного раза.

Идея заключается в том, чтобы перемешивать элементы генеральной совокупности случайным образом и затем отобрать в выборку необходимое количество элементов. Это похоже примерно на то, как тасуют колоду игральных карт, чтобы затем сдать нужное для игры количество карт.

Здесь вновь помогает Excel.

Воспользуемся вновь примером с мотками проволоки. Для этого предпринимаем следующие шаги.

1. Откроем новый Лист 5. В ячейках A1:B1 укажем "Выборка без возврата", а также поместим заголовки ***Номер мотка*** и ***Случайное*** в соответственно в ячейки A3 и B3. Затем в столбец A введем номера элементов исследуемой совокупности под номерами от 1 до 120. Но можно поступить и проще – из предыдущего Листа 4 скопировать весь диапазон A1:A123.

2. Следующий столбец B с помощью генератора случайных чисел заполним равномерно распределенными случайными числами, находящимися в интервале от 0 до 1. С этой целью в столбец B вводим функцию **СЛЧИС**, для чего вписываем формулу **=СЛЧИС()** в ячейку B4 (рис.1.11).

3. Затем дважды щелкаем по маркеру заполнения в правом нижнем углу ячейки B4 (маленький черный крестик) и протягиваем его до ячейки B123. Весь столбец B в диапазоне B4:B123 оказывается заполненным случайными числами. Здесь можно сразу можно установить нужную разрядность, укажем три знака после запятой.

То, что получилось, можно видеть на рис.1.12. Как видно, здесь показаны только начало и конец этой глубокой таблицы.

В нашем случае (что, естественно, совершенно случайно) эти числа расположились в последовательности 0,577; 0,301; 0,567, 0,362 … и т.д. – вплоть до 0,448.

4. Выделим теперь ячейки, содержащие функцию **СЛЧИС** (B4:B123), щелкнем правой кнопкой мыши и выберем ***Копировать*** в контекстном меню.

5. При выделенных ячейках B4:B123 щелкнем правой кнопкой еще раз и укажем в контекстном меню опцию ***Специальная вставка***. В появившемся окне отметим пункты ***Значения*** и ***Нет,*** затем снимем отметки с пунктов ***Пропускать пустые ячейки*** и ***Транспортировать***. После чего ***ОК***.

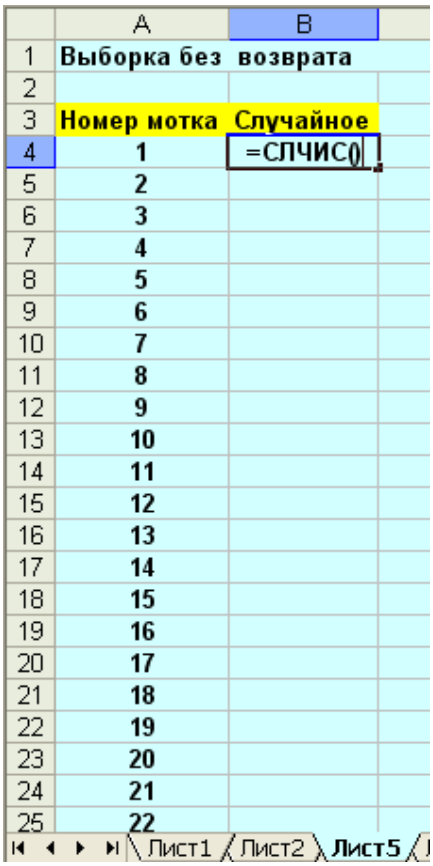


Рис.1.11. Лист Excel с данными для формирования бесповторной выборки

Наш лист теперь будет выглядеть так, как показано на рис. 1.13.

6. Выделим теперь целиком сам массив и случайные числа (A4:B123), т.е. без тех ячеек, где "сидят" заголовки. Выберем ***Сортировка*** в меню ***Данные***. В диалоговом окне ***Сортировка диапазона*** укажем позицию ***Случайное*** в ниспадающем меню списка ***Сортировать по*** и щелкнем по кнопке ***По возрастанию*** (рис.1.14). После чего – клавиша ***ОК***.

	A	B
1	Выборка без	возврата
2		
3	Номер мотка	Случайное
4	1	0,577
5	2	0,301
6	3	0,567
7	4	0,362
8	5	0,509
9	6	0,343
10	7	0,599
11	8	0,334
12	9	0,315
13	10	0,506
14	11	0,404
15	12	0,144
16	13	0,752
17	14	0,113
18	15	0,277
19	16	0,525
20	17	0,706
21	18	0,026
22	19	0,458
23	20	0,286
24	21	0,110

	A	B
100	97	0,268
101	98	0,619
102	99	0,752
103	100	0,531
104	101	0,816
105	102	0,932
106	103	0,679
107	104	0,589
108	105	0,281
109	106	0,416
110	107	0,471
111	108	0,270
112	109	0,083
113	110	0,557
114	111	0,315
115	112	0,107
116	113	0,700
117	114	0,642
118	115	0,924
119	116	0,711
120	117	0,607
121	118	0,676
122	119	0,097
123	120	0,448

Рис.1.12. Случайные числа до сортировки

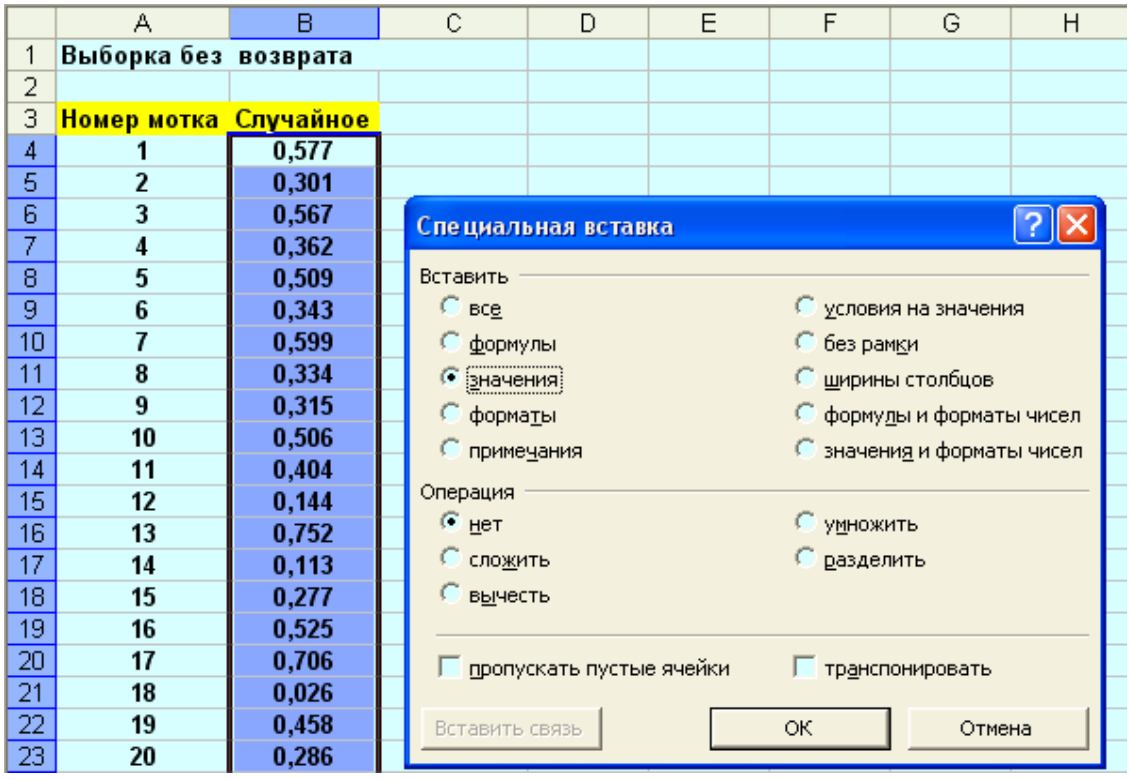


Рис. 1.13. Диалоговое окно *Специальная вставка*

Тем самым будет выполнена сортировка строк на основе тех значений, которые располагаются в столбце со случайными числами.

В итоге указанные манипуляции позволяют отсортировать содержимое обоих столбцов (A и B) таким образом, чтобы обеспечить упорядочение чисел во втором столбце. В результате все элементы генеральной совокупности будут перемешены (перетасованы) случайным образом. Ну, а чтобы осуществить собственно выборку, нужно взять первые ***n*** элементов из этой перемешанной генеральной совокупности.

Окончательный результат представлен на рис.1.15. Как видно, числа первого столбца (номера мотков) расположены в случайном порядке. В зависимости от требуемого объема формируемой случайной выборки следует отсчитать первые ***n*** значений. Так, в нашем случае для получения выборки из 12 единиц следует отобрать следующие мотки: 4, 29, 53, 55, 64, 65, 67, 92, 94, 103 и 110.

Отметим, что полученная таким образом случайная выборка будет обладать теми же свойствами, что и выборка, построенная с использованием таблицы случайных чисел.

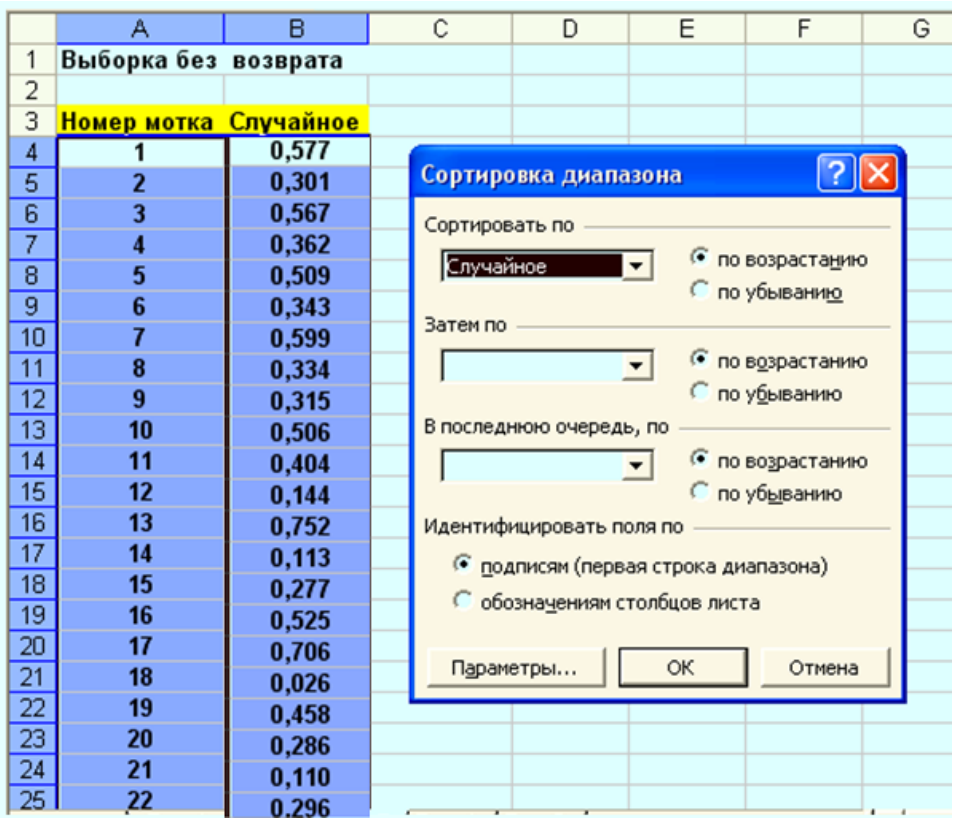


Рис. 1.14. Диалоговое окно *Сортировка диапазона*

	A	B
1	Выборка без возврата	
2		
3	Номер мотка	Случайное
4	94	0,800
5	29	0,524
6	103	0,494
7	67	0,537
8	92	0,246
9	64	0,479
10	110	0,533
11	56	0,616
12	4	0,479
13	55	0,795
14	53	0,784
15	65	0,504
16	99	0,860
17	98	0,825

Рис.1.15. Случайные числа после сортировки

1.12. Обработка экспериментальных результатов

Итак, мы получили представление о следующих полезных вещах:

– какой должен быть по объему выборочный массив;

– каким образом организовать его формирование.

Иными словами, мы можем ответить на вопросы "сколько" и "как".

Теперь нужно выяснить, какие потребуется определить показатели для выборки, которые позволили бы количественно судить о уже самой генеральной совокупности.

1.12.1. Определение среднего арифметического и стандартного отклонения

Удобно знакомиться с необходимыми процедурами, рассматривая конкретный пример. Продолжим знакомую же задачу с мотками стальной проволоки.

Для выбранного массива из 12 мотков стальной проволоки было проведено взвешивание каждого из них и был получен следующий первичный ряд экспериментальных данных, кг:

60,5; 62,8; 58,1; 57,5; 62,4; 61,2; 60,9; 62,2; 58,5; 61,0; 54,2; 58,6.

Следует определить:

– наличие грубого промаха (выскакивающих значений);

– среднее генеральной совокупности (истинное значение) $\tilde{x}$ *путем расчета среднего арифметического выборки* $\bar{x}$ *и доверительного интервала (предельной ошибки)* Δx.

Запустим программу Excel и выделим следующий по номеру рабочий лист (Лист 6).

1. Вначале представим полученные данные в табличной форме в виде двух столбцов (рис.1.16), снабдив их соответствующими заголовками "Номер мотка" (ячейка A1) и "Вес, кг" (ячейка B1). В диапазоне A2:A13 укажем номера мотков от 1 до 12, а в B2:B13 – соответствующие весовые значения. Далее в ячейках A14 и A15 запишем "Ср. ариф-е" и "Ст. откл-е". Зарезервируем дополнительные ячейки B14 и B15, в которых будут размещены рассчитанные значения среднего арифметического $\bar{x}$ и среднеквадратичного отклонения S_n.

2. Выделим ячейку B14 , в которую будет помещен искомый результат; затем активизируем ***Мастер функций*** кнопкой f_x (или же в строке меню используем команды ***Вставка/Функция***).

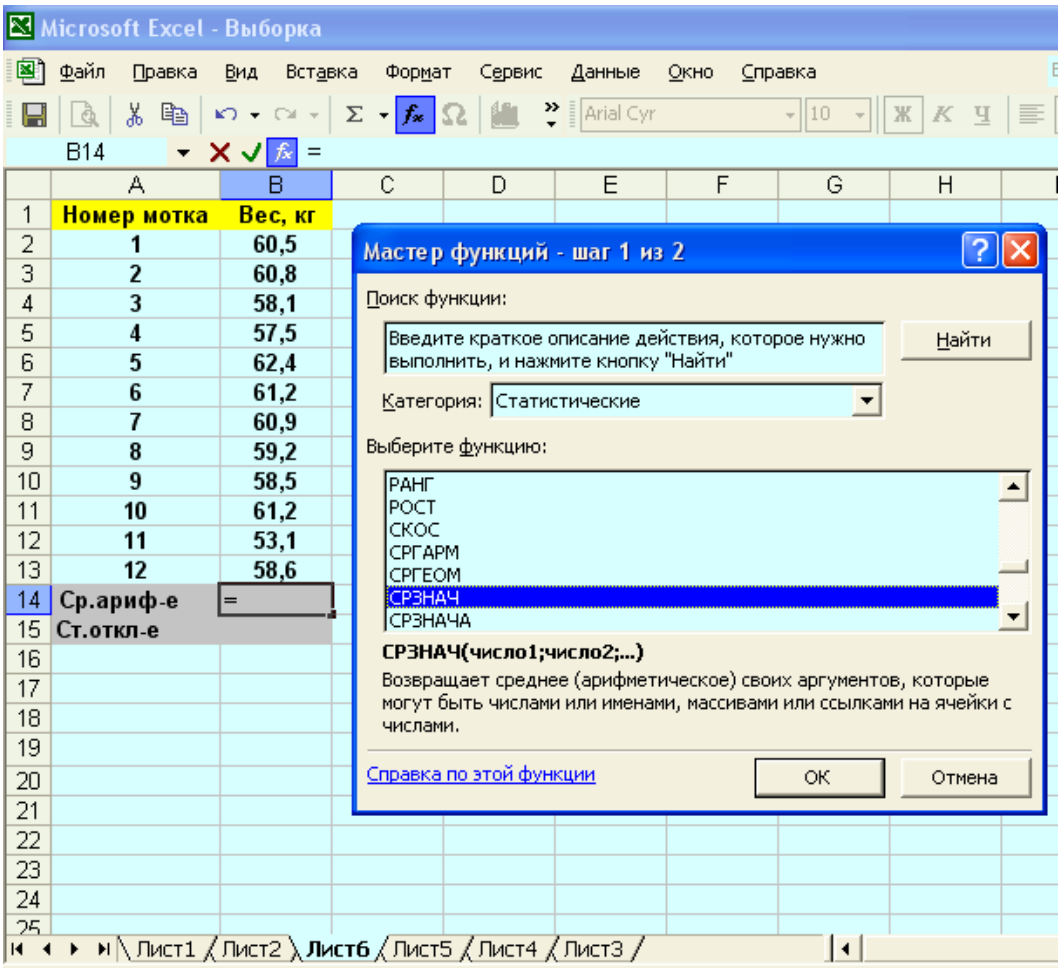


Рис.1.16. Диалоговое окно *Мастер функций*

3. В появившемся диалоговом окне выберем нужную функцию из списка (все функции разбиты на категории); для этого в окне ***Категория*** укажем требуемую опцию под названием ***Статистические***.

Затем в нижнем окне (где перечислены функции) выделим собственно нужную функцию – ***Срзнач*** и нажмем на кнопку ***ОК***.

4. Появится окно ***Аргументы функции***. Подведем маркер к окну ввода ***Число 1*** и выделим все ячейки второго столбца, т.е. это те ячейки, где расположены числовые результаты нашего опыта (B2:B13). После чего нажимаем на кнопку ***ОК***.

Пояснение. Если панель ***Аргументы функции*** закрывает значительную часть поля листа и числа в столбце B не видны, то можно поступить следующим образом. Свернем диалоговое окно, для чего нажмем кнопку справа от поля ввода. В результате можно будет увидеть всю таблицу. Выделим столбец, где "сидят" наши данные, после этого вновь нажмем на кнопку – окно вновь полностью раскроется (рис.1.17).

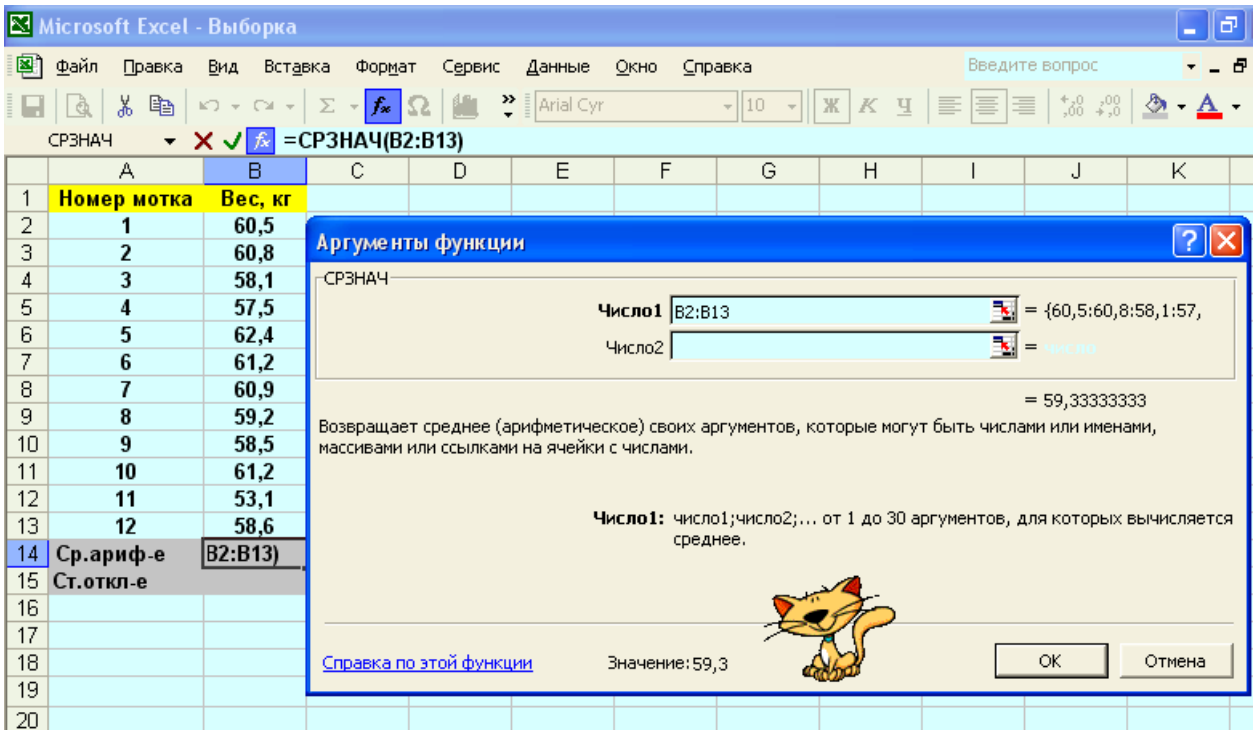


Рис.1.17. Диалоговое окно *Аргументы функции*

5. Подобные манипуляции проделаем и для последней ячейки B15 – среднеквадратичного (стандартного) отклонения. Сделаем только одно

замечание. При работе с ***Мастер функций*** нужно будет активизировать функцию ***Стандотклон***.

В обеих ячейках (B14 и B15) будут размещены рассчитанные значения среднего арифметического $\bar{x}$ и среднеквадратичного отклонения S_n.

Теперь надлежит рассчитать доверительный интервал. Для этого в ячейке A16 запишем "Доверит. интервал", а в ячейке B16 предусмотрим размещение самого результата вычисления.
Последующие манипуляции будут следующими.

6. В диалоговом окне в имеющихся строках ввода укажем последовательно величину уровня значимости ***Альфа*** (0,05), значение стандартного отклонения ***Станд.откл*** (отметим ячейку B15) и объем выборочного массива ***Размер*** (12).

На рис.1.18 показан соответствующий лист Excel.

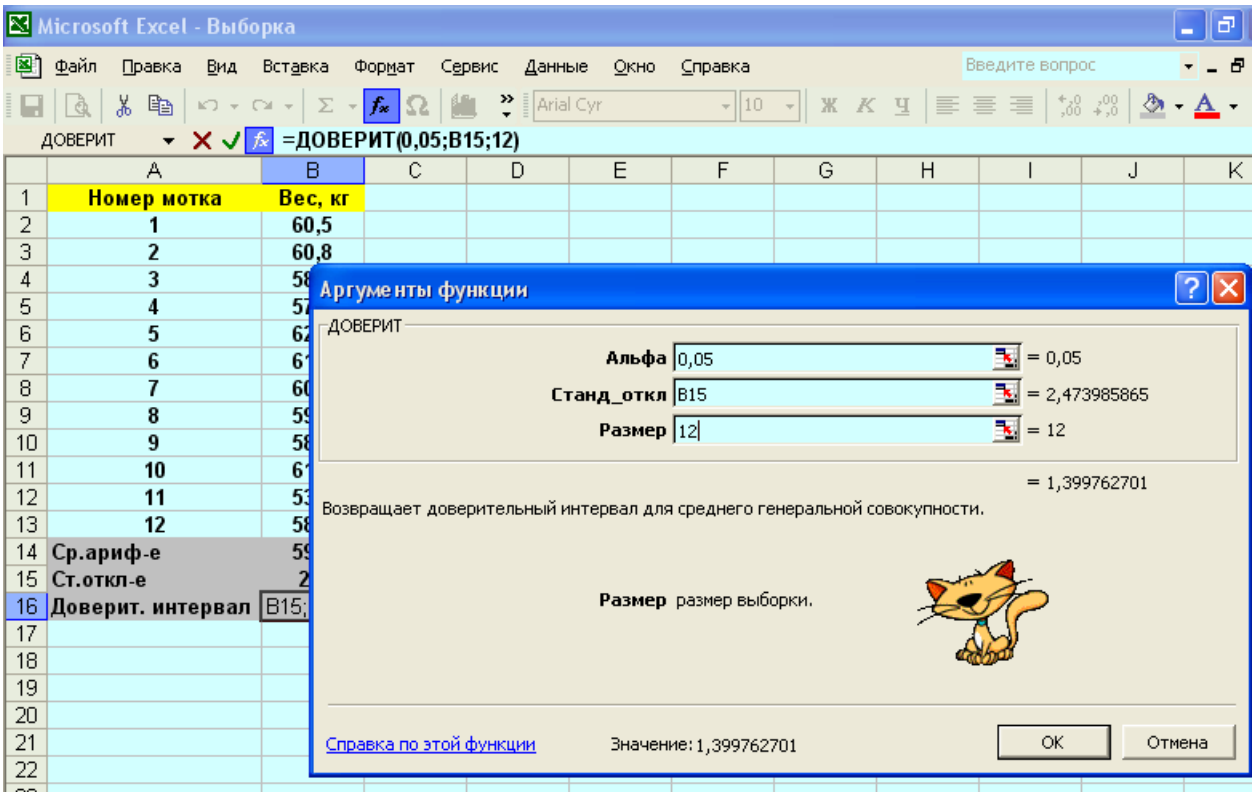


Рис.1.18. Вид диалоговое окна при вычислении доверительного интервала

В окончательном виде наши табличные данные можно видеть на рис.1.19. Заметим, что здесь приведены итоговые значения с учетом необходимой разрядности (с одним знаком после запятой, как и у самих исходных данных).

1.12.2. Нахождение грубого промаха

Одна из обязательных процедур статистической обработки результатов измерений включает оценку так называемых *грубых промахов* или *выскакивающих значений*. Ими могут быть случайные ошибки большой величины, вероятность которых, вообще-то говоря, весьма мала. Однако понятное желание интуитивно выбросить такой выпадающий (как нам кажется) результат

	A	B
1	**Номер мотка**	**Вес, кг**
2	**1**	**60,5**
3	**2**	**60,8**
4	**3**	**58,1**
5	**4**	**57,5**
6	**5**	**62,4**
7	**6**	**61,2**
8	**7**	**60,9**
9	**8**	**59,2**
10	**9**	**58,5**
11	**10**	**61,2**
12	**11**	**53,1**
13	**12**	**58,6**
14	**Ср.ариф-е**	**59,3**
15	**Ст.откл-е**	**2,5**
16	**Доверит. интервал**	**1,4**
17		

Рис.1.19. Итоговые данные расчета показателей выборки

и получить радующий глаз практически ровный ряд измерений является совершенно необоснованным и порочным. Ибо таким способом легко получить завышенную и совершенно фиктивную точность измерений. Задача статистического анализа и состоит как раз в том, чтобы "подозрительный" результат подвергнуть специальной проверке, на основании чего и принимается окончательное решение о его судьбе – сохранить или выкинуть (с облегчением) из массива.

В теории статистики известны различные рекомендации по поводу отсева грубых промахов. Один из таких способов – широко применяемый *метод максимального относительного отклонения*. Его принято использовать в случае малой выборки ($n \le 25$). Кстати сказать, число производимых измерений обыкновенно и бывает относительно небольшим и редко превышает 10-20 замеров.

Для расчета максимального относительного отклонения $\tau_{макс}$ часто используется соотношение, которое оценивает относительное отличие проверяемого (или так называемого крайнего) результата $x_{кр}$ от среднего арифметического $\bar{x}$ выраженное в долях среднеквадратичной ошибки S_n. Полученный результат (его абсолютное значение) затем сравнивается со специальным статистическим эталоном. Математическая статистика как раз и заботится о создании таких эталонов, которые называются *критическими* или *табличными* значениями. Сама процедура сопоставления вычисленной характеристики с табличным значением именуется *проверкой гипотезы* или *проверкой на адекватность*.

Итак, используется следующее соотношение:

$$\tau_{макс} = \left| \frac{\bar{x} - x_{кр}}{s_n} \right| \le \tau_{табл} \quad .$$

В случае весьма малой выборки ($n \le 10$) принято использовать уточненное выражение для определения $\tau_{макс}$:

$$\tau_{макс} = \left| \frac{1}{\sqrt{(n-1)/n}} \frac{\bar{x} - x_{кр}}{s_n} \right| \le \tau_{табл} \quad .$$

Таким образом, для выявления выскакивающих значений нужно рассчитанное значение $\tau_{макс}$ сопоставить с табличным $\tau_{табл}$, т.е. проверить соотношение $\tau_{макс} \le \tau_{табл}$.

В том случае если это соотношение соблюдается, то проверяемый результат считается входящим в данную числовую совокупность и его отбрасывать нельзя. В случае же обратного итога, т.е. $\tau_{макс} > \tau_{табл}$, анализируемый результат признается ошибочным и его надлежит исключить из дальнейшего рассмотрения.

Заметим, что при статистических расчетах такую проверку принято осуществлять для заданной доверительной вероятности P или же соответствующего уровня значимости $\alpha = P - 1$. Как ранее отмечалось, опыт использо-

вания статистики в коммерческой практике показывает, что приемлемым значением α является 0,05 (5%-ный *уровень риска*).

Вернемся теперь к нашему примеру.

Чтобы намеренно обострить ситуацию, предположим, что один из мотков оказался по весу за пределами оговоренных условий поставки. Полагаем, что этого в действительности не случилось, но нам нужен просто пример для анализа.

Итак, как видно, в полученных измерениях явно настораживает 11-й результат, равный 53,1. При расчете максимального относительного отклонения воспользуемся выражением для $n > 10$.

Вновь обратимся к Excel. Действуем следующим образом.

7. На листе 6 в ячейках A18 и A19 последовательно запишем "Крайнее значение" и "Макс.отн.откл-е", а в соседнюю ячейку B18 потом поместим $x_{кр}$=53,1.

8. Выделим ячейку B19 и в поле ввода формул запишем **=(B14–B18)/B15**. После этого нажмем клавишу Enter и в ячейке появится (после округлений) число 2,52 (рис.1.20).

Полученный результат $\tau_{макс}$ теперь надлежит сопоставить с табличным значением $\tau_{табл}$. В имеющемся *Приложении 2* для заданных условий (n = 12 и α=0,05) находим, что $\tau_{табл}$= 2,39. Легко видеть, что выполняется условие $\tau_{макс} > \tau_{табл}$, поэтому с вероятностью 95% (или иначе с риском ошибиться на 5%) можно утверждать, что проверяемый результат является грубым промахом и его следует удалить из данного массива.

	A	B
1	Номер мотка	Вес, кг
2	1	60,5
3	2	60,8
4	3	58,1
5	4	57,5
6	5	62,4
7	6	61,2
8	7	60,9
9	8	59,2
10	9	58,5
11	10	61,2
12	11	53,1
13	12	58,6
14	Ср.ариф-е	59,3
15	Ст.откл-е	2,5
16	Доверит. интервал	1,4
17		
18	Крайнее значение	53,1
19	Макс. отн. откл-е	2,52

Рис.1.20. Результаты расчета максимального относительного отклонения

Теперь надлежит осуществить процедуру изъятия "нехорошего" результата из табличного массива. Для этого вновь воспользуемся замечательными возможностями Excel.

9. В таблице удалим результат, соответствующий номеру 11. Для этого выделим ячейку B12 и нажмем клавишу ***Delete***. После этого в ячейках, указывающих значения среднего арифметического, стандартного и доверительного интервала отклонения, автоматически устанавливаются их обновленные показатели (рис.1.21).

Формально процедуру отсева полагается повторять и для следующего крайнего значения. Однако предварительно следует пересчитать $\bar{x}$ и S_n для выборки нового объема, т.е. уже для $n - 1$. Такой пересчет Excel, как мы видим, выполнил (рис.1.21). Попробуем выполнить аналогичную проверку для следующего по ранжиру крайнего значения. Им является результат под номером 4, равный 57,5. Для этого случая рассчитанная величина $\tau_{макс}$ составляет 1,52 (рис.1.22). Значение $\tau_{табл}$, извлеченное из *Приложения 2*, в этом случае составит 2,34, т.е. выполняется соотношение $\tau_{макс} \leq \tau_{табл}$, Следовательно, этот результат входит в данную совокупность (с вероятностью 95%).

	A	B
1	Номер мотка	Вес, кг
2	1	60,5
3	2	60,8
4	3	58,1
5	4	57,5
6	5	62,4
7	6	61,2
8	7	60,9
9	8	59,2
10	9	58,5
11	10	61,2
12	11	
13	12	58,6
14	Ср.ариф-е	59,9
15	Ст.откл-е	1,6
16	Доверит. интервал	0,9

Рис.1.21. Показатели выборочного массива после удаления грубого промаха

На этом наш анализ закончен. В окончательном виде результат измерения среднего генеральной совокупности (истинного значения) $\tilde{x}$ можно представить так:

$$\tilde{x} = \bar{x} \pm \Delta x = 59{,}9 \pm 0{,}9 \; кг.$$

Следовательно, если исключить из рассмотрения такой прискорбный результат, как 53,1 кг (напомним, мы его намеренно придумали), то в целом отдел снабжения фирмы, занятой производством весьма полезных метизов

(гвоздей и шурупов), может быть вполне удовлетворен аккуратностью и добросовестностью своего коммерческого партнера. Ибо с надежностью 95% метизно-металлургический завод поставляет продукцию по весу, практически совпадающему с заявленным параметром (при среднем весе, равном 59,9 кг, и регламентированном показателе в 60,0 кг), и при этом со значительно меньшей погрешностью (± 0,9 кг), чем это предусмотрено условиями контракта (± 3,0 кг). Результат просто блестящий!

	A	B
1	Номер мотка	Вес, кг
2	1	60,5
3	2	60,8
4	3	58,1
5	4	57,5
6	5	62,4
7	6	61,2
8	7	60,9
9	8	59,2
10	9	58,5
11	10	61,2
12	11	
13	12	58,6
14	Ср.ариф-е	59,9
15	Ст.откл-е	1,6
16	Доверит. интервал	0,9
17		
18	Крайнее значение	57,5
19	Макс. отн. откл-е	1,52

Рис.1.22. Проверка на грубый промах следующего крайнего значения

1.13. Построение гистограмм

Одним из способов графического изображения результатов статистического распределения какой-либо величины x по количественному признаку является представление их в виде гистограмм или столбчатых диаграмм. Гистограмма распределения позволяет оценить, сколько раз измеренные значения x укладываются в заданные дискретные промежутки $\Delta_1,\ldots, \Delta_k$ (интервалы или разряды), охватывающие весь диапазон изменения этой величины. Гистограмма графически строится в виде столбцов, образующих совокупность смежных прямоугольников, построенных на прямой линии. Их высота (по оси ординат) соответствует количеству попаданий чисел из рассматриваемого массива n в заданный интервал изменения x, на который опирается столбик (на горизонтальной оси).

Гистограммы обычно строят для абсолютных частот (это когда считают число попаданий f_k в k-м разряде). Иногда удобнее анализировать относительные частоты w_k (частости), которые определяются как $w_k = f_k/n$. Здесь $n=\Sigma f_k$, т.е. сумма отдельных частот f_k дает общее количество измерений n, т.е. объем выборки.

Целесообразность подобного графического изображения полученных экспериментальных результатов представляется разумной в тех случаях, когда приходится исследовать большой массив однородных случайных величин, подверженных очевидному статистическому разбросу. В этом отноше-

нии типичная ситуация, с которой приходится сталкиваться, скажем, коммерсанту, – нужно проанализировать характер распределения по количественному признаку каких-либо характеристик товарной продукции, численности персонала фирмы, затрат на какую-то коммерческую деятельность и проч.

Обычно строить гистограммы имеет резон в тех случаях, когда рассматривается массив из достаточно большого числа измерений ***n***. Считается, что такие построения представляются более надежными для ***n***>75-100, а при ***n***<25-30 использование гистограмм в статистическом смысле становится неоправданным.

Как мы знаем, замечательный Excel превосходно справляется со всякого рода статистическими головоломками и может, ко всему прочему, успешно строить гистограммы. Для этого приложение оснащено специальной программой ***Гистограмма***, входящей в особый пакет ***Анализ данных***.

Познакомимся с приемами построения гистограмм с помощью Excel. Для этого рассмотрим достаточно типичный для коммерческой практики пример.

Среди туристических фирм Уральского региона, занятых организацией заграничных поездок на отдых, было проведено исследование по поводу их финансовых затрат на проведение рекламных компаний. Обследованию были подвергнуты 100 наиболее успешных туристических заведений.

Результаты статистического наблюдения приведены в табл.1.4, где перечислены турфирмы и указаны их средние месячные затраты на рекламу (тыс. руб). Необходимо полученные данные обследования представить в графической форме в виде гистограммы распределения.

Итак, исходная информация представлена в табл.4, в которой курсивом выделены порядковые номера турфирм (колонки серого цвета), а в соседних столбцах указаны сами результаты финансовых трат.

А теперь обратимся к вновь к Excel. Действуем в следующей последовательности.

1. Запустим Excel и откроем рабочий лист, которому затем можно придумать подходящее наименование. Присвоим нашему документу незамысловатое название "Гистограмма".
2. Теперь введем исходные данные. Для этого в ячейку A1 поместим надпись "Номера турфирм", а затем одним столбиком, набравшись

терпения, разместим в ячейках A2:A101 сами экспериментальные результаты (их, как мы помним, 100!).

Пояснение. Попытка представить эти данные в виде компактной таблицы (с несколькими колонками) приведет к тому, что Excel станет рассматривать каждый столбец как самостоятельную совокупность чисел и выдаст потом результаты для каждой колонки отдельно. Получится очевидная нелепость.

3. В ячейке B1 запишем "Затраты на рекламу, тыс. руб" и введем наши исходные данные.

Таблица 1.4

Исходные данные для построения гистограммы

1	25	*26*	29	*51*	96	*76*	95
2	22	*27*	64	*52*	20	*77*	30
3	38	*28*	46	*53*	21	*78*	50
4	91	*29*	31	*54*	53	*79*	41
5	54	*30*	28	*55*	48	*80*	75
6	52	*31*	65	*56*	60	*81*	49
7	30	*32*	52	*57*	73	*82*	67
8	69	*33*	41	*58*	44	*83*	21
9	62	*34*	27	*59*	84	*84*	35
10	88	*35*	38	*60*	103	*85*	24
11	59	*36*	48	*61*	24	*86*	39
12	97	*37*	57	*62*	75	*87*	56
13	101	*38*	34	*63*	68	*88*	42
14	59	*39*	38	*64*	52	*89*	74
15	22	*40*	55	*65*	46	*90*	29
16	75	*41*	44	*66*	37	*91*	84
17	84	*42*	51	*67*	98	*92*	57
18	82	*43*	46	*68*	57	*93*	83
19	35	*44*	57	*69*	24	*94*	28
20	38	*45*	26	*70*	39	*95*	37
21	66	*46*	73	*71*	41	*96*	48
22	71	*47*	51	*72*	57	*97*	51
23	38	*48*	49	*73*	62	*98*	60
24	90	*49*	41	*74*	55	*99*	70
25	57	*50*	56	*75*	41	*100*	92

Для рассматриваемого массива чисел нам нужно получить в виде сводной таблицы основные статистические характеристики. С этой целью воспользуемся специальной программой ***Описательная статистика***.

4. Для того чтобы ее запустить, в главном меню выберем последовательно пункты ***Сервис/Анализ данных/Описательная статистика***, после чего щелкнем по кнопке ***ОК***.

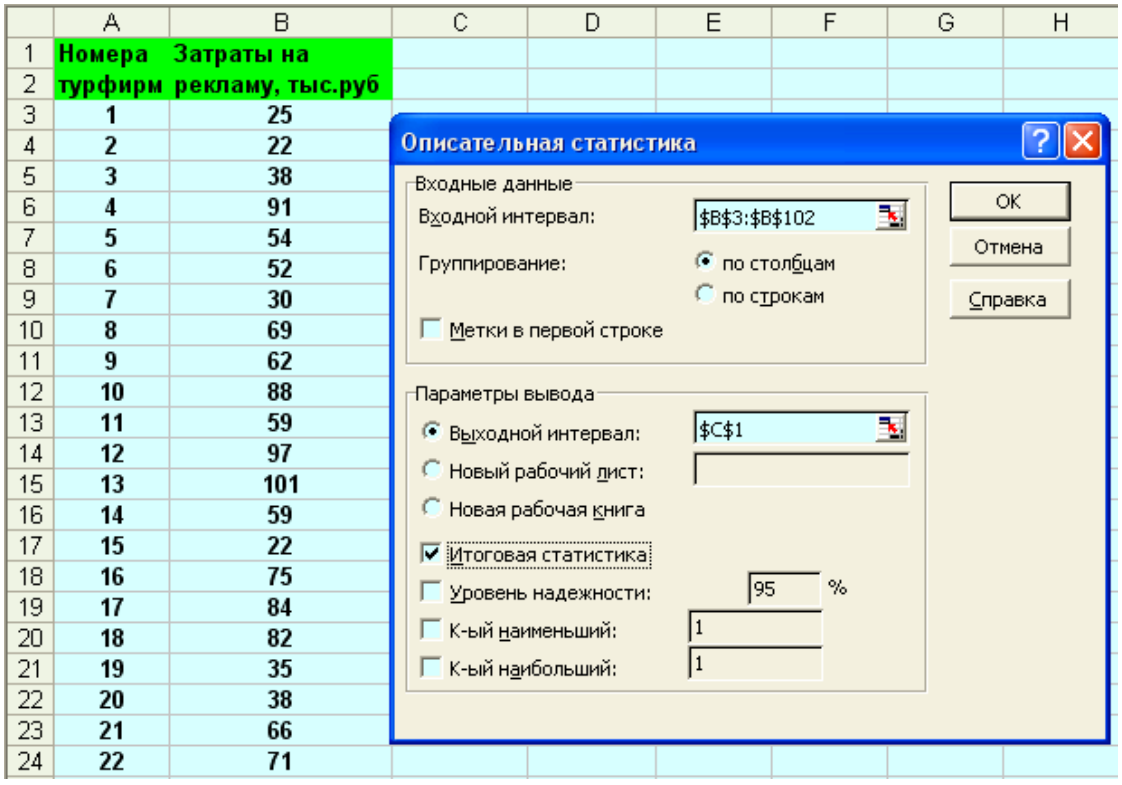
5. Далее заполним появившееся диалоговое окно ввода данных и параметров вывода, для этого проделаем следующее (рис.1.23):

– укажем ***Входной интервал*** (в виде абсолютных ссылок $B$3:$B$101);

– отметим способ ***Группирования*** (в нашем случае по столбцам);

– выделим ***Выходной интервал***, для этого достаточно указать левую верхнюю ячейку будущего диапазона; пусть это будет ячейка $C$1;

– установим флажок, показывающий, что нам нужна информация в виде ***Итоговой статистики***, после чего – кнопка ***ОК***.



	A	B
1	Номера	Затраты на
2	турфирм	рекламу, тыс.руб
3	1	25
4	2	22
5	3	38
6	4	91
7	5	54
8	6	52
9	7	30
10	8	69
11	9	62
12	10	88
13	11	59
14	12	97
15	13	101
16	14	59
17	15	22
18	16	75
19	17	84
20	18	82
21	19	35
22	20	38
23	21	66
24	22	71

Рис.1.23. Диалоговое окно ввода параметров *Описательная статистика*

Результаты вычисления Excel представит нам в табличной форме (рис.1.24). При этом таблица содержит шапку "Показатели статистики" (в ячейках C1:C2) и "Результат расчета" (D1:D2). В данном случае Excel к указанной текстовой манипуляции не причастен – это наше деяние. Дело в том, что при построении таблицы Excel автоматически ввел заголовок "Столбец 1", который помещался в ячейке C1, поэтому приходится менять надпись на более приемлемую. Для этого дважды щелкнем последовательно в этих ячейках, чтобы они превратилась в поля ввода; запишем нужные слова и затем нажмем клавишу ***ENTER***.

Отметим еще одну особенность. В столбце C содержатся наименования статистических характеристик. Целиком текст может и не помещаться, так как ширина ячейки оказывается недостаточной. Чтобы избавиться от такого очевидного неудобства, проделаем следующее. Ставим курсор на правой границе заголовка столбца C и дважды щелкаем левой клавишей. Результат налицо – ширина столбца стала такой, что весь текст благополучно влез в его поле и можно читать самые длинные названия.

	A	B	C	D
1	Номера	Затраты на	Показатели	Результат
2	турфирм	рекламу, тыс.руб	статистики	расчета
3	1	25	Среднее	54
4	2	32	Стандартная ошибка	2
5	3	38	Медиана	52
6	4	91	Мода	57
7	5	54	Стандартное отклонение	21
8	6	52	Дисперсия выборки	445
9	7	30	Эксцесс	0
10	8	69	Асимметричность	1
11	9	62	Интервал	83
12	10	88	Минимум	20
13	11	59	Максимум	103
14	12	97	Сумма	5394
15	13	101	Счет	100
16	14	59		
17	15	39		

Рис.1.24. Результаты расчета статистических показателей

Сделаем некоторые пояснения по поводу содержимого листа Excel, приведенного на рис.1.24. В столбце C перечислены статистические характеристики, а в соседнем столбце D указаны их значения. Как видно, усердный Excel выдал большой набор разных статистических параметров. Они, конечно, полезны и способны дать всестороннюю статистическую оценку рассматриваемого массива. Однако нам нужна и другая информация – необходимо "нарисовать картинку", т.е. в виде графика получить наглядное представле-

ние о том, каков характер распределения рекламных затрат по количественному признаку. В связи с этим перейдем к следующему этапу нашей работы с Excel – он должен помочь нам построить искомую гистограмму распределения.

Теперь-то собственно и обратимся к программе ***Гистограмма***. Сначала нам нужно задать разряды (интервалы), на которые исследуемый массив следует разделить. Вообще-то говоря, эту процедуру можно и не выполнять – сообразительный Excel сам автоматически выберет разряды, равномерно распределив их между минимальным и максимальным числовыми значениями. Однако попробуем осуществить разбиение на интервалы самостоятельно. В нашем случае удобно задаться шагом, равным 10, так как измеренные величины располагаются между 20 и 103 (см. рис.1.24). Выберем ячейку E1 и там запишем "Разряды". Затем, начиная с ячейки E3, вводим столбиком выбранные границы разрядов – 10, 20, 30 и т.д. вплоть до 120. Здесь мы намеренно выберем крайние интервалы (1-10 и 111-120), в которых заведомо содержатся нулевые результаты.

В итоге придуманные нами разряды будут размещены в ячейках E3:E14.

6. В главном меню отыщем опцию ***Сервис***, а далее ***Анализ данных/Гистограмма*** (рис.1.25). В появившемся диалоговом окне заполним ***Входные данные***.

– укажем ***Входной интервал***, отмечая диапазон ячеек, где располагается наш массив, а именно: $B$3:$B$102 (называем только те ячейки, в которых сидят числа);

– следующая позиция – ***Интервал карманов*** (в терминологии Excel под карманами понимаются те разряды, на которые нужно будет разделить рассматриваемый массив); отметим знакомые уже нам ячейки $E$3:$E$14 (с числами);

– проигнорируем флажок ***Метки***, тем самым Excel должен будет самостоятельно позаботиться о заголовке.

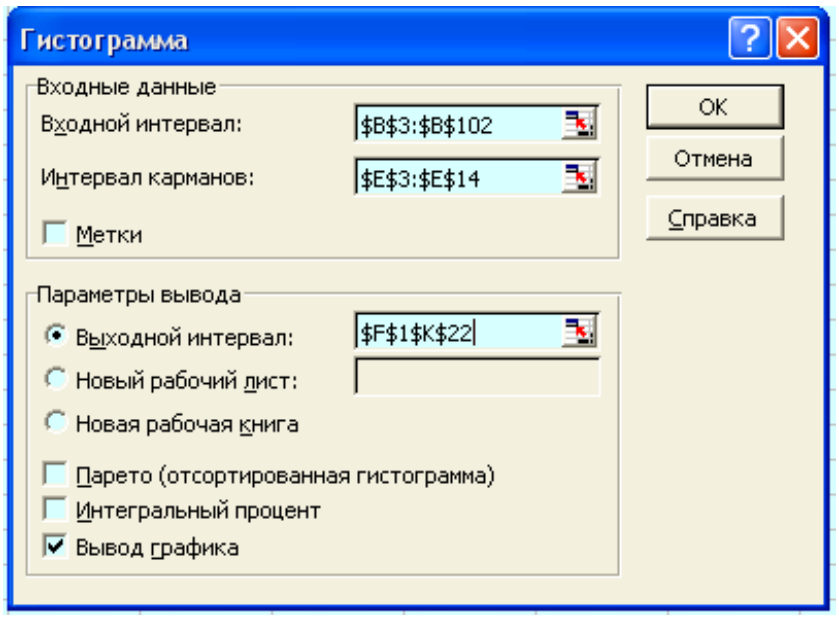


Рис.1.25. Диалоговое окно ввода параметров *Гистограмма*

7. Теперь заполним ***Параметры вывода***.

– укажем ***Выходной интервал***; здесь достаточно дать ссылку на левую верхнюю ячейку выходного интервала, поскольку размер выходного диапазона будет создан автоматически; Пусть такой ячейкой станет ячейка с координатами $F$1.

– отметим флажком ***Вывод графика***, после чего нажмем на кнопку ***ОК.***

Что же в итоге получим? Трудолюбивый Excel выдаст искомые материалы – табличную форму распределения чисел рассматриваемого массива и саму гистограмму (рис.1.26).

Займемся сначала изучением таблицы, содержащейся на рис.1.26. В ней содержатся две колонки (F и G), указывающие заданные интервалы (под названием "Карман") и количество попаданий в каждый интервал чисел данного массива (именуется "Частота"). Причем эти заголовки были внесены самим Excel.

	E	F	G
1	Разряды	Карман	Частота
2		10	0
3	10	20	1
4	20	30	14
5	30	40	15
6	40	50	17
7	50	60	22
8	60	70	9
9	70	80	7
10	80	90	7
11	90	100	6
12	100	110	2
13	110	120	0
14	120	Еще	0

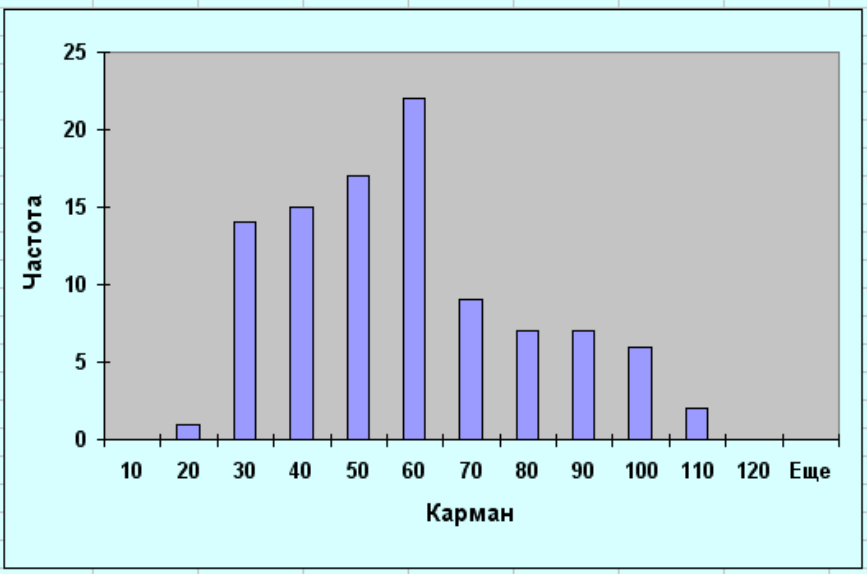


Рис.1.26. Табличная и графическая формы представления результатов

В последней строке стоит разряд с забавным именем "Еще". Он показывает интервалы (возможные), которые располагаются за пределами выбранного нами диапазона. Поскольку эта строка никакой полезной информации в данном случае не несет, то удалим из обеих колонок содержимое последних строк (F14 и G14).

Перейдем теперь к гистограмме. Кстати, в первоначальном виде Excel ее изобразил в весьма сжатом виде (рис.1.26), поэтому после нехитрых манипуляций она станет выглядеть так, как это показано на рис.1.27.

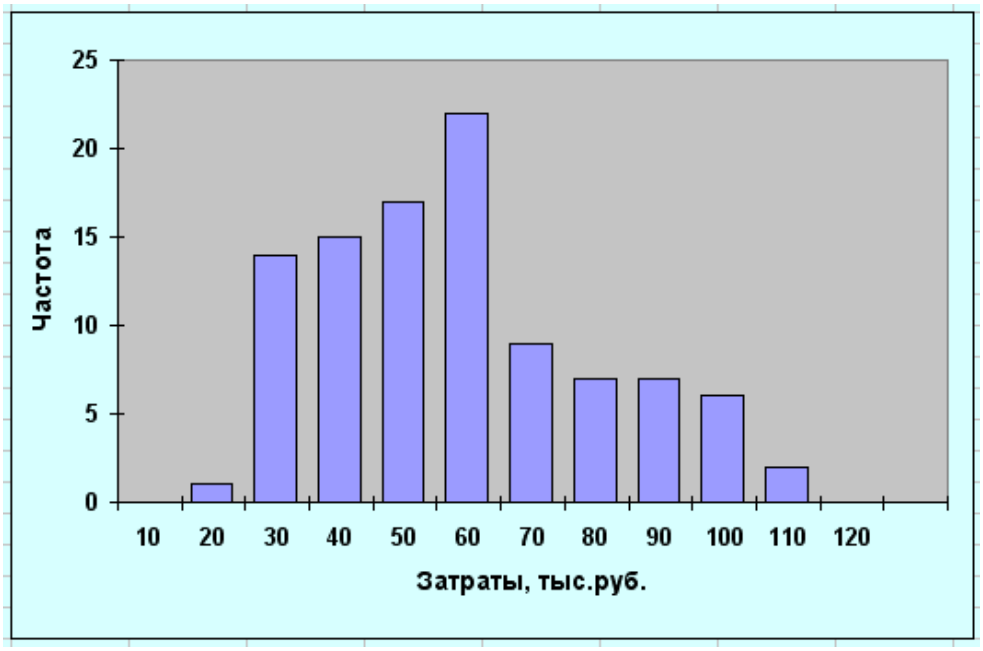


Рис.1.27. Окончательный вид гистограммы

И, наконец, последнее. Эти же результаты, которые мы изобразили гистограммой, можно представить и в другой форме – в виде *кривой частоты* или так называемого *многоугольника* (*полигона*) *распределения*. В этом случае график будет представлять собой ломаную линию, построенную на основании расчета среднеинтервальных значений рассматриваемого массива (эту работу, понятное дело, исполнит вновь неутомимый Excel).

Делается это следующим образом.

8. Поместим курсор в поле диаграммы, чтобы высветилась надпись ***Область построения диаграммы***.

9. Щелкнем правой клавишей – появится контекстное меню. В нем выберем опцию ***Тип диаграммы,*** а затем на вкладке ***Стандартные*** в категории ***Тип*** укажем ***График***. После чего – клавиша ***ОК***.

На рис.1.28 можно будет полюбоваться нашим новым компьютерным произведением.

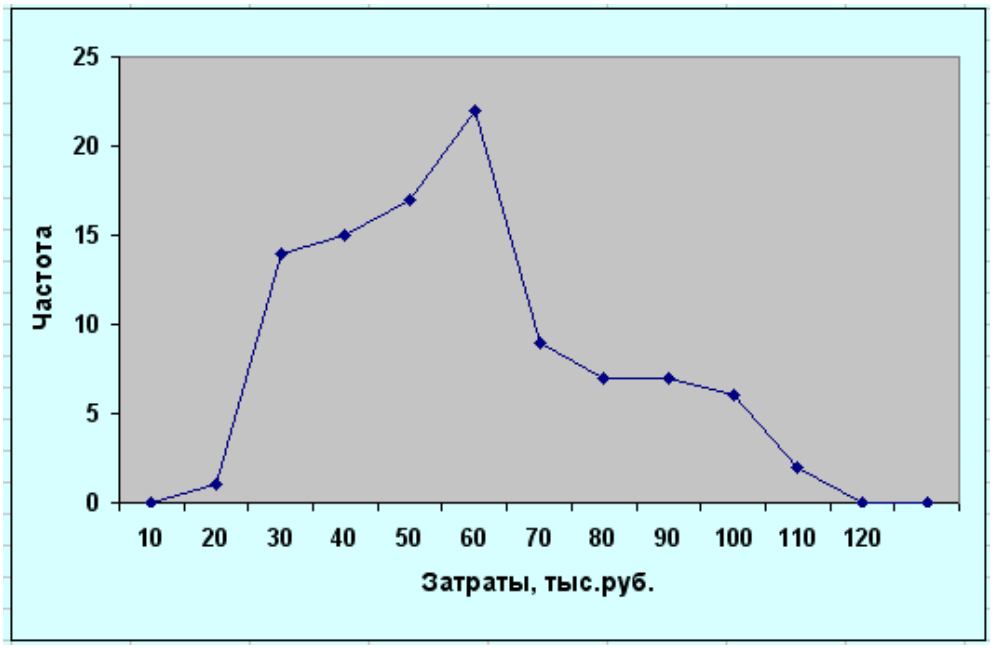


Рис.1.28. Полигон распределения финансовых затрат

На этом нашу задачу можно считать решенной.

2. КОРРЕЛЯЦИОННАЯ СВЯЗЬ И ЕЕ СТАТИСТИЧЕСКОЕ ИЗУЧЕНИЕ В КОММЕРЧЕСКОЙ ДЕЯТЕЛЬНОСТИ

Качество корреляционной зависимости обратно пропорционально плотности точек

Один из постулатов Мэрфи

Исследование отдельных статистических объектов позволяет получить о них вполне полезную информацию и описать их стандартными показателями. При этом изучаемую совокупность можно представить в виде ряда распределения путем ранжирования (в порядке возрастания или убывания анализируемого количественного признака), дать характеристику этой совокупности, указав центральные значения ряда (среднее арифметическое, медиану, моду), размах варьирования, форму кривой распределения. Такого рода сведения могут быть вполне достаточными в случаях, когда приходится иметь дело с *одномерными* данными (т.е. лишь с *одной* характеристикой, например, зарплатой) о каждой единице совокупности (скажем, о сотруднике фирмы).

Когда же мы анализируем *двумерные* данные (например, зарплата и образование), всегда есть возможность изучать каждое измерение по отдельности – как часть одномерной совокупности данных. Однако реальную отдачу можно получить лишь при совместном изучении обоих параметров. Основной смысл такого подхода – создается возможность выявить *взаимосвязь* между ними.

2.1. Типы зависимостей

Следовательно, помимо традиционных измерений и последующих вычислений при анализе статистических данных приходится решать проблему и более высокого уровня – выявление функциональной зависимости между *воздействующим фактором* и *регистрируемой* (изучаемой) *величиной*. Указанные ситуации весьма типичны в статистической практике, и в этом смысле аналитическая работа коммерсанта весьма богата такими примерами.

Зависимость одной случайной величины от значений, которые принимает другая случайная величина, в статистике называется *регрессией*. Если этой зависимости придан аналитический вид, то такую форму представления изображают *уравнением регрессии*.

Процедура поиска предполагаемой зависимости между различными числовыми совокупностями обычно включает следующие этапы:

– установление значимости связи между ними*;

– возможность представления этой зависимости в форме математического выражения (уравнения регрессии).

Первый этап в указанном статистическом анализе касается выявления так называемой *корреляции* или *корреляционной зависимости*. Корреляция рассматривается как признак, указывающий на *взаимосвязь* ряда числовых последовательностей. Иначе говоря, корреляция характеризует *силу взаимосвязи* в данных. Если это касается взаимосвязи двух числовых массивов x_i и y_i, то такую корреляцию называют *парной*.

При поиске корреляционной зависимости обычно выявляется вероятная связь одной измеренной величины x (для какого-то ограниченного диапазона ее изменения, например от x_1 до x_n) с другой измеренной величиной y (также изменяющейся в каком-то интервале $y_1 \ldots y_n$). В таком случае мы будем иметь дело с двумя числовыми последовательностями, между которыми и надлежит установить наличие статистической (корреляционной) связи. На этом этапе пока *не* ставится задача определить, является ли одна из этих случайных величин *функцией*, а другая – *аргументом*. Отыскание количественной зависимости между ними в форме конкретного аналитического выражения $y=f(x)$ – это задача уже другого анализа, регрессионного.

Таким образом, *корреляционный анализ* позволяет сделать вывод о силе взаимосвязи между парами данных x и y, а *регрессионный анализ* используется для *прогнозирования* одной переменной (y) на основании другой (x). Иными словами, в этом случае пытаются выявить причинно-следственную связь между анализируемыми совокупностями.

Схематическое изображений изложенных соображений представлено на рис.2.1.

* Статистический смысл термина *значимость* означает, что анализируемая зависимость проявляется сильнее, чем это можно было бы ожидать от чистой случайности.

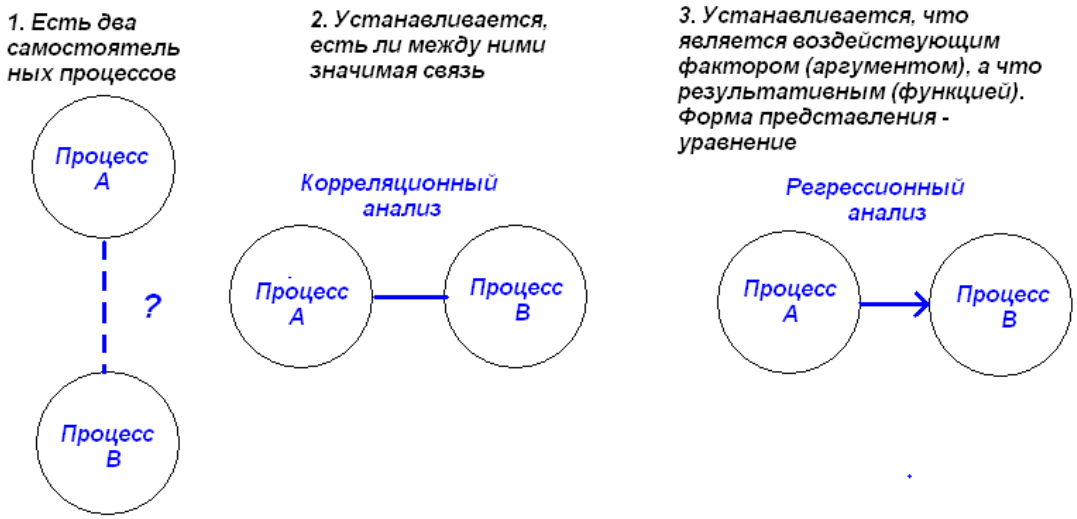


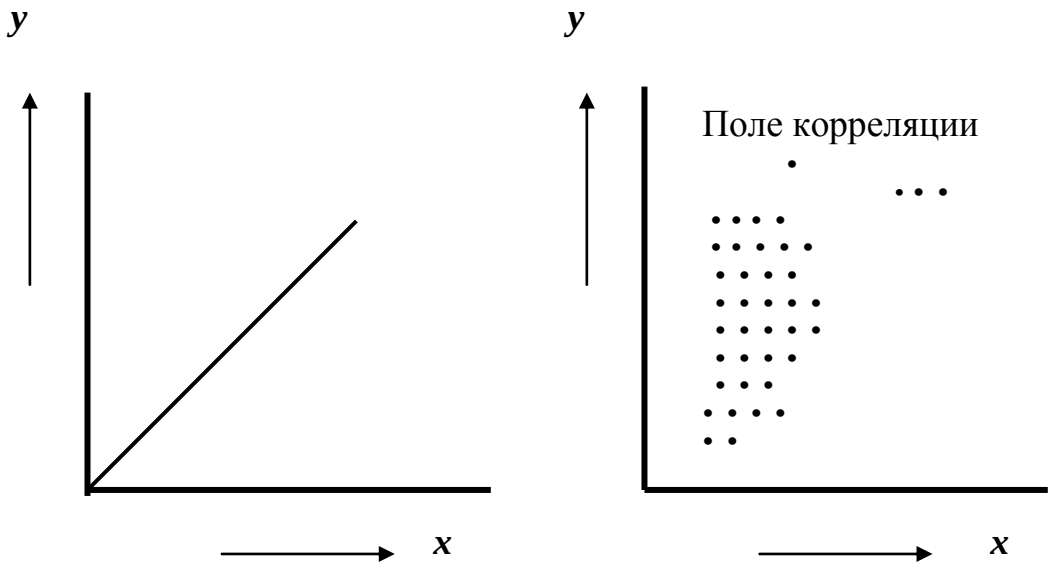
Рис.2.1. Схематическое пояснение сути корреляционного и регрессионного анализов

Строго говоря, принято различать два вида связи между числовыми совокупностями – это может быть *функциональная* зависимость или же *статистическая* (случайная). При наличии функциональной связи каждому значению воздействующего фактора (аргумента) соответствует строго определенная величина другого показателя (функции), т.е. изменение результативного признака всецело обусловлено действием факторного признака.

Аналитически функциональная зависимость представляется в следующем виде:

$$y=f(x).$$

Графически это (при наличии линейной зависимости) может быть представлено в виде прямой линии (рис.2.2*а*).



а б

Рис.2.2. Зависимость функциональная (*а*) и статистическая (*б*)

В случае статистической связи значению одного фактора соответствует какое-то приближенное значение исследуемого параметра, его точная величина является непредсказуемой и поэтому получаемые показатели оказываются случайными величинами. Это значит, что изменение результативного признака ***y*** обусловлено влиянием факторного признака ***x*** лишь частично, т.к. возможно воздействие и иных факторов, вклад которых обозначен как **ε**:

$$y=\varphi(x)+\varepsilon.$$

По своему характеру корреляционные связи – это соотносительные связи. Примером корреляционной связи показателей коммерческой деятельности является, например, зависимость сумм издержек обращения от объёма товарооборота. В этой связи помимо факторного признака ***x*** (объёма товарооборота) на результативный признак ***y*** (сумму издержек обращения) влияют и другие факторы, в том числе и неучтенные, порождающие вклад **ε**.

Такая зависимость графически изображается в виде экспериментальных точек, образующих *поле рассеяния* или, как принято говорить, *поле корреляции* (рис.2.2*б*). Следовательно, такие двумерные данные можно анализировать с использованием *диаграммы рассеяния* в координатах "***x*** - ***y***", которая дает визуальное представление о взаимосвязи исследуемых совокупностей.

Для количественной оценки существования связи между изучаемыми совокупностями случайных величин используется специальный статистический показатель – *коэффициент корреляции* ***r***. Если предполагается, что эту связь можно описать линейным уравнением типа $y=a+bx$ (где ***a*** и ***b*** – константы), то принято говорить о существовании линейной корреляции.

Коэффициент ***r*** – это безразмерная величина, она может меняться от 0 до ±1. Чем ближе значение коэффициента к единице (неважно, с каким знаком), тем с большей уверенностью можно утверждать, что между двумя рассматриваемыми совокупностями переменных существует линейная связь. Иными словами, значение какой-то одной из этих случайных величин (***y***) существенным образом зависит от того, какое значение принимает другая (***x***).

Если окажется, что $r=1$ (или -1), то имеет место классический случай чисто функциональной зависимости (т.е. реализуется идеальная взаимосвязь).

При анализе двумерной диаграммы рассеяния можно обнаружить различные взаимосвязи. Простейшим вариантом является линейное соотношение, которое выражается в том, что точки размещаются случайным образом вдоль прямой линии. Диаграмма свидетельствует об отсутствии взаимосвязи, если точки расположены случайно и при перемещении слева направо невозможно обнаружить какой-либо уклон (ни вверх, ни вниз).

Если точки на ней группируются вдоль *кривой* линии, то диаграмма рассеяния характеризуется *нелинейной взаимосвязью*. Такие ситуации вполне возможны. Тем не менее, для удобства понимания сути корреляционного соотношения мы ограничимся рассмотрением варианта линейной зависимости.

2.2. Методы определения корреляционной связи

Корреляцию и регрессию принято рассматривать как совокупный процесс статистического исследования и поэтому их использование в статистике часто именуют *корреляционно-регрессионным анализом*.

Если между парами совокупностей связь просматривается вполне очевидная (скажем, ранее нами это исследовалась, есть публикации на данную тему и т.д.), то, минуя стадию корреляции, можно сразу приступать к поиску уравнения регрессии.

Если же исследования касаются какого-то нового процесса, ранее не изучавшегося, то наличие связи между совокупностями является предметом специального поиска.

При этом условно можно выделить методы, которые позволяют оценить *качественно* наличие связи, и методы, дающие *количественные* оценки.

Чтобы выявить наличие *качественной* корреляционной связи между двумя исследуемыми числовыми наборами экспериментальных данных, существуют различные методы, которые принято называть *элементарными*.

Ими могут быть приемы, основанные на следующих операциях:

– *параллельном сопоставлении* рядов;

– построении *корреляционной или групповой таблиц*;

– *графическом изображении* с помощью *поля корреляции.*

Другой метод, более сложный и статистически надежный, – это *количественная* оценка связи посредством расчета коэффициента корреляции и его статистической проверки.

Познакомимся со способом оценки корреляционной связи посредством расчета коэффициента корреляции, рассмотрев конкретный пример.

2.3. Расчет коэффициента парной корреляции и его статистическая проверка

Существуют различные аналитические приемы определения коэффициента *r*. Наиболее часто рекомендуется использовать выражение:

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

где x_i и y_i – текущие значения единиц обеих совокупностей, а n – число измерений (элементов) в каждой совокупности.

Зная коэффициент корреляции, можно дать качественно-количественную оценку тесноты связи. Используются, например, специальные табличные соотношения (так называемая шкала Чеддока). Её представление может иметь следующий вид (табл. 2.1):

Т а б л и ц а 2.1

Качественная оценка тесноты связи

Величина коэффициента парной корреляции	Характеристика силы связи
До 0,3	Практически отсутствует
0,3–0,5	Слабая
0,5–07	Заметная
0,7–0,9	Сильная
0,9–0,99	Очень сильная

Такие оценки носят общий характер и не претендуют на статистическую строгость, поскольку не дают гарантий на вероятностную достоверность.

Поэтому в статистике принято использовать более надежные критерии для оценки степени тесноты связи, основываясь на рассчитанных значениях *коэффициента парной корреляции (КПК*).

Здесь может помочь только эталон, с которым можно было бы сравнить вычисленную характеристику. Статистика как раз и занимается созданием таких эталонов, которые называются *критическими* или *табличными значениями.*

Процедуру установления корреляционной зависимости принято называть *проверкой гипотезы.* Ее принято проводить в следующей последовательности:

– вычисление линейного коэффициента парной корреляции между совокупностями случайных величин x_i и y_i;

– его статистическая оценка (проверка значимости).

Статистическую оценку КПК проводят путем сравнения его абсолютной величины с табличным (или критическим) показателем $r_{крит}$, значения которого отыскиваются из специальной таблицы.

Если окажется, что $|r_{расч} \geq r_{крит}|$, то с заданной степенью вероятности (обычно 95%) можно утверждать, что между рассматриваемыми числовыми совокупностями существует значимая линейная связь. Или по-другому – гипотеза о значимости линейной связи не отвергается.

В случае же обратного соотношения, т.е. при $|r_{расч} < r_{крит}|$, делается заключение об отсутствии значимой связи.

Перейдем теперь к рассмотрению конкретного примера. Рассмотрим несколько шутливую ситуацию с привлечением известных героев популярного мультфильма "Трое из Простоквашино".

Дядя Федор с озабоченностью отметил, что в продолжение прошедшей недели у кота Матроскина заметно снизилась эффективность ловли мышей. Сам Матроскин объяснил означенный настораживающий факт тем, что погода в это время портилась и средняя температура имела тенденцию к устойчивому понижению. Однако пес Шарик посчитал, что причина совершенно в ином – просто Матроскин разленился, стал много больше спать, и мышам стало вольготнее.

Дядя Федор решил внимательно проанализировать возникшую проблему и собрал необходимые для этого данные за $n=7$ *дней. Полученные результаты он аккуратно свел в табл.2.2, где указал число пойманных мышей за каждый день исследуемой недели, среднюю дневную температура за этот период и, наконец, число часов, которые кот отвел себе для сна.*

На основании этих данных дяде Федору важно было выяснить, есть ли корреляция между названными показателями и какая из возможных причин – изменение температуры или продолжительность сна – сказались в большей степени на результативности поимки серых грызунов.

Т а б л и ц а 2.2

Снижение эффективности мышиной охоты кота Матроскина и ее возможные причины

Дни	Число пойманных	Средняя дневная	Продолжительность

	мышей	температура, °С	сна, часы
1	7	17	7
2	8	15	8
3	5	13	8
4	6	12	10
5	5	12	11
6	4	10	10
7	3	8	12

Работать будем с приложением Excel, поэтому запустим его:

1. Нажмем кнопку ***Пуск*** в панели задач (находится слева на самой нижней полосе ***Рабочего стола***), а затем откроем в всплывающем меню опцию ***Программы***.

2. Выберем пункт ***Microsoft Excel***; откроется книга Excel с указанием рабочего листа 1 (внизу экрана будет высвечен знак ***Лист 1***).

Подготовим табл.2.2 в виде четырех столбцов. Вначале заготовим "шапку" таблицы. Для этого в ячейках A2; B2; C2 и D2 запишем соответственно "Дни", "Число пойманных мышей", "Средняя дневная температура, °С" и "Длительность сна, часы". Затем разместим сами числовые наборы соответственно в диапазонах ячеек A3:A9, B3:B9, C3:C9 и D3:D9 (рис.2.3).

Укажем также таблицу, в которой поместим расчетные значения коэффициента. Выделим для этого диапазон ячеек C13:D16, где будут находиться необходимые заголовки. Сами же значения коэффициента корреляции будем помещать в ячейки D15 и D16.

Далее определим коэффициент корреляции с помощью ***Мастера функций***. Вначале выполним расчет для соотношения "Количество пойманных мышей – средняя дневная температура".

	A	B	C	D	E
1					
2	Дни	Число пойманных мышей	Средняя дневная температура,°С	Длительность сна, часы	
3	1	7	17	7	
4	2	8	15	8	
5	3	5	13	8	
6	4	6	12	10	
7	5	5	12	11	
8	6	4	10	10	
9	7	3	8	12	
10					
11					
12					
13			Причина	Коэффициент корреляции	
14			Температура		
15			Сон		
16					
17					

Рис.2.3. Исходные данные и расчет коэффициента корреляции

Действуем в такой последовательности:

3. В итоговой таблице активизируем ячейку D15, куда и будет помещено первое рассчитанное значение КПК.
4. Запустим ***Мастер функций*** (ищем в инструментальной строке значок f_x) и в появившемся диалоговом окне укажем требуемую категорию – ***Статистические***, а затем выделим нужную функцию ***Коррел***, а затем – ***ОК*** (рис.2.4).
5. В появившейся панели ***Коррел*** нужно заполнить текстовые поля для ***Массив 1*** (т.е. указать диапазон ячеек B3:B9) и для ***Массив 2*** (C3:C9); для этого выделим в нашей таблице последовательно 2-ю и 3-ю колонки (там, напомним, размещены числовые значения мышей и температуры), причем каждый раз в соответствующих окнах должен находиться маркер (мерцающая вертикальная черточка); выделенная колонка по периметру будет обрамлена бегущей пунктирной линией (рис.2.5); после чего кнопка ***ОК***.

Аналогичным образом поступим для расчета второго коэффициента, используя вновь 2-ю колонку, а также следующую 4-ю колонку ("Длительность сна, часы").

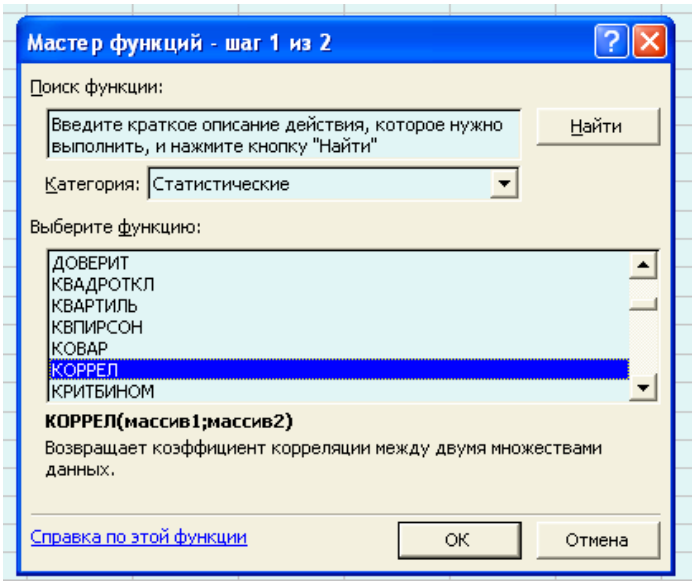


Рис.2.4. Диалоговое окно *Мастер функций*

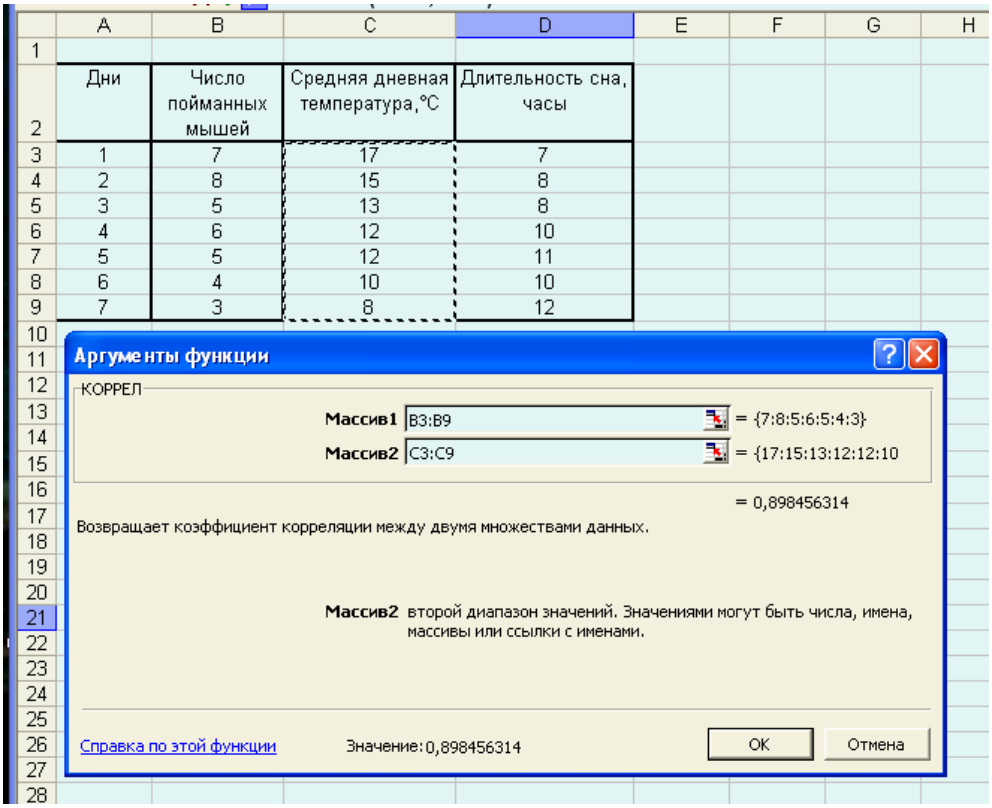


Рис.2.5. Диалоговое окно ввода параметров корреляции

В выделенных ячейках D15 и D16 (рис.2.6) появятся числа, указывающие соответствующие значения коэффициентов корреляции. После установления нужной разрядности в окончательном виде получим следующие значения: $r_{расч1}$ =0,898 и $r_{расч2}$ = –0,764.

Первый коэффициент показывает, насколько заметна теснота связи параметров "Количество пойманных мышей – средняя дневная температура". Второй показатель характеризует другую изучаемую связь "Количество пойманных мышей – длительность сна, часы". Отметим, что второй коэффициент имеет знак минус, что говорит об обратном соотношении указанных параметров (в общем-то, понятно, чем больше спит Матроскин, тем менее эффективной становится охота на мышей).

	A	B	C	D	E
1					
2	Дни	Число пойманных мышей	Средняя дневная температура,°С	Длительность сна, часы	
3	1	7	17	7	
4	2	8	15	8	
5	3	5	13	8	
6	4	6	12	10	
7	5	5	12	11	
8	6	4	10	10	
9	7	3	8	12	
10					
11					
12					
13			Причина	Коэффициент корреляции	
14					
15			Температура	0,898	
16			Сон	-0,764	
17					

Рис. 2.6. Рассчитанные значения коэффициента корреляции

Теперь надлежит дать статистическую оценку выполненных нами расчетов, т.е. проверить на адекватность рассматриваемые события. Для этого сопоставим расчетные значения коэффициентов $r_{расч}$ с табличным показателем $r_{крит}$. Используя *Приложение 3*, находим, что для уровня значимости (т.е. вероятности допустимой ошибки в прогнозе) α=0,05 и заданного числа измерений n табличное значение $r_{крит}$=0,754.

Как видно, в обоих случаях выполняется соотношение $|r_{расч} \geq r_{крит}|$, а посему озабоченный дядя Федор с уверенностью 95% может полагать, что между рассматриваемыми числовыми совокупностями существует корреляционная связь. Вместе с тем резонно утверждать, что обсуждаемые причины вполне можно ранжировать по степени влияния – более существенную роль играют погодные условия, но и мнение пса Шарика, как видно, имеет статистическое обоснование.

Пояснение. Заметим, что в таблице для $r_{крит}$ (*Приложение 3*) вместо привычных значений числа измерений n стоит показатель f, характеризующий *степень свободы*. Число степеней свободы, как отмечалось, определяется путем вычисления разности между количеством опытов (измерений) n и числом коэффициентов (констант), которые уже рассчитаны по результатам этих опытов, т.е. $f=n-k$, где k – это количество вычисленных констант. В нашем случае в формуле для r участвуют две константы $\bar{x}$ и $\bar{y}$, поэтому на r остается только n–2 "свободных" измерений, т.е. $n-2=7-2=5$.

2.4. О ложной корреляции (влияние "третьего фактора")

Часто корреляцию и причинную обусловленность считают синонимами. Этот тезис имеет определенные основания, поскольку если н*е*что является причиной чего-либо другого, то можно говорить о связи первого и второго и, следовательно, об их коррелированности (например, действие и результат, проверка и качество, капиталовложения и прибыль, окружающая среда и степень комфортности).

Однако корреляция может быть и без причинной обусловленности. Это можно представить так: корреляция – лишь число, которое указывает на то, что б*о*льшим значениям одной переменной соответствуют б*о*льшие (или же м*е*ньшие) значения другой переменной. Корреляция *не* может объяснить, *почему* эти две переменные связаны между собой. Так, корреляция не объясняет, почему капиталовложения порождают прибыль (или наоборот). Корреляция просто констатирует, что между этими величинами существует определенное соответствие.

Одним из возможных оснований для существования "корреляции без причинной обусловленности" является наличие некоторого скрытого, ненаблюдаемого, *третьего фактора*, который "маскируется" под другую переменную. В результате фиксируется так называемая "*ложная корреляция*".

Допустим, нами выявлена высокая корреляция между приемом на работу новых менеджеров и созданием новых производственных мощностей. Возможно, именно менеджеры являются "причиной" капиталовложений в новые производственные мощности? Или же, наоборот, создание новых производственных мощностей послужило "причиной" приема на работу новых менеджеров? Скорее всего, однако, здесь проявляется действие третьего фактора – высокая потребность в продукции фирмы, что и послужила причиной и приема на работу новых менеджеров и создания новых производственных мощностей.

В истории статистики известен один классический по этому поводу пример. Он касается курьезного исследования под условным названием "Аисты приносят детей". Так, в шведской столице в течение 73 лет регистрировалось число новорожденных в год (y) и число аистов (x), которых содержало население. Указанные данные были сведены в таблицу и по ним был рассчитан КПК. Он оказался близок к единице, так что формально никакой статистики и не требовалось для проверки. Все экспериментальные точки аккуратно улеглись на прямую, т.е. практически указанную связь следовало бы толковать как чисто функциональную.

Поскольку утверждение, содержащее в упомянутом тезисе, довольно сомнительное, было решено поискать другое разумное объяснение. Оказалось, что одновременные синхронные изменения числа аистов и числа детишек объясняются изменением среднего уровня жизни жителей Стокгольма. Эта переменная первоначально не являлась предметом рассмотрения, отчего и случился такой забавный курьез вследствие ложной корреляции.

В качестве статистического показателя может быть использован также *коэффициент (индекс) детерминации* (*причинности*) $\boldsymbol{R^2}$, который равен квадрату коэффициента корреляции. Он показывает, в какой мере изменчивость $\boldsymbol{y}$ (результативного признака) объясняется поведением $\boldsymbol{x}$ (факторного признака), или иначе: какая часть общей изменчивости $\boldsymbol{y}$ вызвана собственно влиянием $\boldsymbol{x}$. Этот показатель вычисляется путём простого возведения в квадрат коэффициента корреляции. Тем самым доля изменчивости $\boldsymbol{y}$, определяемая выражением $1-\boldsymbol{R^2}$, оказывается необъясненной.

Допустим к примеру, что коэффициент корреляции совокупности данных, относящихся к производственным затратам, равняется 0,869. Следовательно, значение $\boldsymbol{R^2}$ равно

$$\boldsymbol{R^2}= 0{,}869^2 = 0{,}756 \text{ или } 75{,}6\%.$$

Это значение $\boldsymbol{R^2}$ говорит о том, что 75,6% вариации (изменчивости), скажем, недельных затрат объясняется количеством изделий, выпущенных за неделю. Остальная часть (24,4%) вариации общих затрат объясняется какими-то другими причинами. Это значит, что более чем на 75% мы знаем, что влияет на изменение изучаемого параметра, но почти на 25% ничего путного сказать не можем о причинах наблюдаемой изменчивости.

Величина этого коэффициента меняется в пределах от 0 до 1. Чем ближе он к единице, тем, следовательно, меньше в нашей модели процесса влияние неучтенных факторов и тем больше оснований считать, что указанная зависимость отражает степень эффективности воздействия изучаемого фактора.

2.5. Измерение степени тесноты связи между качественными признаками (ранговая корреляция)

При определении корреляционной зависимости нужно было иметь числовой набор двух совокупностей. Однако возможны случаи, когда имеющиеся данные *не* поддаются выражению числом единиц.

Это обстоятельство заставляет прибегать к использованию так называемых *непараметрических методов*. Они позволяют измерять интенсивность взаимосвязи между качественными (атрибутивными) признаками. В основу непараметрических методов положен принцип нумерации значений статистического ряда. Каждой единице массива присваивается порядковый номер (ранг) в ряду, который будет упорядочен (ранжирован) по уровню признака.

Следовательно, важным условием является возможность сделать рассматриваемые совокупности *упорядоченными*.

Предварительное представление о наличии или отсутствии связи между рассматриваемыми массивами можно получить, если сопоставить последовательность взаимного расположения рангов факторного (воздействующего) и результативного (подверженного влиянию) признаков. Для этого ранги измеренных значений факторного признака располагают в порядке возрастания. Если ранги результативного признака обнаруживают тенденцию к увеличению, то можно говорить о наличии прямой связи. Если картина противоположная, то и связь толкуется как обратная.

В статистике известны коэффициенты корреляции, основанные на использовании рангов. Одним из таковых является *коэффициент корреляции рангов Спирмена*. Он основан на рассмотрении *разности рангов* значений *факторного* и *результативного* признаков и ее обозначают как $\boldsymbol{d_i}$.

Представим себе, что имеются две выборки, которые классифицированы по каким-то двум признакам: $\boldsymbol{x}$ и $\boldsymbol{y}$.

Выборки (их объем): 1, 2, 3, …, $\boldsymbol{n}$

1-я совокупность (признак $\boldsymbol{x}$): $\boldsymbol{x_1, x_2}$, $\boldsymbol{x_3}$, …, $\boldsymbol{x_n}$

2-я совокупность (признак $\boldsymbol{y}$): $\boldsymbol{y_1, y_2, y_3}$, …, $\boldsymbol{y_n}$.

Здесь оба параметра $\boldsymbol{x}$ и $\boldsymbol{y}$ принимают только целочисленные значения в количестве, равном $\boldsymbol{n}$.

Тогда формула коэффициента корреляции рангов Спирмена (этот коэффициент обозначают как $\boldsymbol{p}$) имеет следующий вид:

$$p = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$ где $\boldsymbol{d_i = x_i - y_i}$.

Рассмотрим определение этого коэффициента на следующем примере.

Было организовано соревнование по компьютерному программированию среди студенческих команд нескольких университетов. Количество таких команд равнялось 12. На предварительном этапе экспертная группа дала прогнозную оценку ожидаемых результатов конкурса, представив ее в виде ранжированного ряда (в порядке убывания). В основу подобного анализа экспертами были учтены разные факторы: уровень профессиональной подготовки команд, их прошлое участие в аналогичных соревнованиях и, соответственно, имеющиеся достижения, наличие научной школы и известные традиции в области программирования. После завершения соревнования были получены фактические данные, характеризующие распределение мест среди команд (их ранжированное положение). В табл. 2.3 приведены соответствующие данные: прогнозные оценки экспертов и итоговые результаты.

Итак, в табл.2.3 укажем условные порядковые номера команд, их место (ранг), которое было предсказано экспертами, а также действительные результаты в виде баллов, набранных командами (по пятибалльной шкале), и фактические места, ими полученные.

Т а б л и ц а 2.3

Расчетная таблица для определения коэффициента ранговой корреляции

Порядковый номер команд	Ранг команд по результатам оценки экспертов R_x	Итоговые баллы команд по результатам соревнования	Ранг команд по результатам соревнования R_y
1	*2*	*3*	*4*
1	6	3,3	3
2	5	3,0	6
3	11	2,8	7
4	4	4,1	2
5	8	2,1	12
6	3	2,7	8
7	10	2,5	10
8	12	2,3	11
9	7	3,2	4

10	9	2,6	9
11	1	3,1	5
12	2	4,5	1

Как видно из результатов сопоставления прогнозных и действительных рангов, общая картина выглядит достаточно пестрой. В одних случаях ранги были вполне совпадающими (например, у команд под номерами 2, 8 и 12, но особенно полное совпадение у команд с номерами 7 и 10), у других же заметно различались (например, у команд под номерами 3, 5, 6 и 11). Возникает вопрос: насколько точно результаты экспертной оценки (прогноза) предугадали действительные итоги соревнования по программированию?

Задачу решим, используя компьютерные расчеты.

1. Итак, запустим программу Excel. В открывшемся рабочем листе Excel (***Лист 1***) сформируем таблицу, в которой поместим данные, соответствующие содержимому столбцов 1-4 из табл.2.3. Кроме того, добавим столбцы E и F, в которых поместим соответственно значения разности рангов $\boldsymbol{d=R_x - R_y}$ и $\boldsymbol{d^2}$. Укажем также итоговую строку 14. В результате таблица будет располагаться в ячейках A1:F14.

2. Заполним столбец E. Для этого в ячейке E2 запишем =B2–D2, появится цифра 3. Выделим эту ячейку и, когда будет зафиксирован справа крестик (***Маркер заполнения***), протянем его вдоль всей колонки до ячейки E13 – получим полностью сформированный столбец E2:E13. Схожим образом организуем следующий столбец, для чего в ячейке F2 запишем =E2^2. Активизировав эту ячейку F2 (там будет сидеть цифра 9), аналогичным приемом заполним столбец F2:F13. В итоговой ячейке F14 укажем сумму (нужно будет выделить весь столбец F2:F13 и в панели инструментов активизировать опцию ***Автосумма***).

В окончательном виде наша таблица примет следующий вид (рис.2.7).

	A	B	C	D	E	F
1	Порядковый номер команды	Ранг команды по результатам оценки экспертов Rx	Итоговые баллы команд по результатам соревнования	Ранг команды по результатам соревновыания Ry	Разность рангов d=Rx-Ry	d^2
2	1	6	3,3	3	3	9
3	2	5	3	6	-1	1
4	3	11	2,8	7	4	16
5	4	4	4,1	2	2	4
6	5	8	2,1	12	-4	16
7	6	3	2,7	8	-5	25
8	7	10	2,5	10	0	0
9	8	12	2,3	11	1	1
10	9	7	3,2	4	3	9
11	10	9	2,6	9	0	0
12	11	1	3,1	5	-4	16
13	12	2	4,5	1	1	1
14	Итого					98

Рис.2.7. Исходные данные и результаты вспомогательных расчетов

3. Для выполнения последующих расчетов используем итоговый результат, отражающий сумму разностей квадратов рангов, равную 98. Для этого ниже имеющейся таблицы в соответствующих ячейках укажем значение Σd^2=98, размер выборки n=12, а также предусмотрим в ней ячейку, где поместим затем рассчитанное значение коэффициента ранговой корреляции p (ячейка D19).

4. Поместим курсор в ячейку D19, а затем в поле формулы запишем уравнение, по которому будем рассчитать коэффициент p. Выглядит оно так:

=1– 6*(D17)/(D18*(D18^2–1)).

В ячейке появится искомый результат – коэффициент корреляции рангов составляет 0,657. В окончательном виде лист Excel будет иметь вид, представленный на рис. 2.8.

	A	B	C	D	E	F
1	Порядковый номер команды	Ранг команды по результатам оценки экспертов Rx	Итоговые баллы команд по результатам соревнования	Ранг команды по результатам соревновыания Ry	Разность рангов d=Rx-Ry	d^2
2	1	6	3,3	3	3	9
3	2	5	3	6	-1	1
4	3	11	2,8	7	4	16
5	4	4	4,1	2	2	4
6	5	8	2,1	12	-4	16
7	6	3	2,7	8	-5	25
8	7	10	2,5	10	0	0
9	8	12	2,3	11	1	1
10	9	7	3,2	4	3	9
11	10	9	2,6	9	0	0
12	11	1	3,1	5	-4	16
13	12	2	4,5	1	1	1
14	Итого					98
15						
16						
17			d^2 =	98		
18			n =	12		
19			p =	0,657		

Рис. 2.8. Фрагмент рабочего листа Excel с обобщенной таблицей и данными для расчета коэффициента корреляции Спирмена

Как и линейный коэффициент корреляции, коэффициент ранговой корреляции может также меняться от –1 до +1. Если воспользоваться шкалой Чеддока, то по результатам расчета коэффициента p можно предположить наличие заметной прямой зависимости между данными прогноза и фактическими результатами. Однако следует учесть, что ранговый показатель был рассчитан по небольшому объёму исходной информации (n=12). Не является ли отличие рангового коэффициента от нуля лишь результатом случайных совпадений оценок экспертов с результатами конкурса по данным малого числа участвующих команд?

Чтобы ответить на этот вопрос более определенно, оценим статистическую значимость расчетного коэффициента. Для этого его значение $p_{расч}$ нужно сопоставить с критическими (табличными) показателями $p_{табл}$. Используется таблица, напоминающая таблицу t-критерия (*Приложение 4*).

Найдем табличное значение коэффициента $p_{табл}$, для α=0,05 и n=12 его величина составит 0,580. Поскольку $p_{расч} > p_{табл}$ (0,657 и 0,580), то с вероятностью 95% можно утверждать, что исследуемая связь является значимой. Однако для уровня значимости α=0,01 табличное значение $p_{табл}$=0,723. Тем самым уже для вероятности 99% наличие связи становится неочевидной.

Таким образом, общий вывод можно свести к следующему тезису: следовало бы повысить число участвующих команд (увеличить объём выборки), а при отсутствии такой возможности высказанные экспертные оценки следует воспринимать с определенной осторожностью.

Заметим, что коэффициент ранговый корреляции может быть использован не только для оценки связи качественных признаков, но и количественных. Принципиальное условие – значения признаков поддаются ранжированию (как именно – по степени убывания или возрастания – это не столь важно).

3. РЕГРЕССИОННЫЙ МЕТОД ОЦЕНКИ КОММЕРЧЕСКОЙ ДЕЯТЕЛЬНОСТИ

> Мы не можем понять эту формулу, и мы не знаем, что она значит, но мы доказали ее и поэтому знаем, что она должна быть достоверной.
>
> (*Некий профессор математики об одной из теорем Л.Эйлера*)

В практике статистического исследования весьма часто возникает необходимость определить не только корреляционное соотношение между изучаемыми характеристиками, но и установить определенную обусловленность между ними, представив выявленную связь в строгой аналитической форме. В этом случае результат исследования – экспериментальная зависимость воздействия какого-либо фактора (скажем, производительности труда, уровня образования, практического стажа работы и т.д.) на изменение изучаемого параметра (например, величины прибыли фирмы) – может быть не только представлен в виде графика (что весьма наглядно), но и описан математически с использованием аппроксимирующего выражения (эмпирической формулы).

Исследование такой ситуации и является задачей *регрессионного* анализа, который дает *предсказание (прогнозирование)* одной переменной на основании другой. Регрессионный анализ четко распределяет роли между изучаемыми характеристиками – что является *аргументом*, а что *функцией*.

Переменная, которая прогнозируется (функция), обозначается как ***y***, а переменная, которая используется для такого прогнозирования (аргумент или фактор) – это ***x***.

Таким образом, в случае выявления корреляции дается попытка ответить на вопрос: "Существует ли связь?" Целью регрессионного анализа является поиск ответа на уже более сложный вопрос: "Каков вид этой связи? Что на что влияет?" Однако в последнем случае речь не идет о выяснении механизма причинности обнаруженной связи, т.е. не ставится вопрос "Почему существует связь?" Это уже считается проблемой специального исследования, касающегося выявления физической (или социальной) природы изучаемого процесса.

3.1. Аппроксимационные модели

При изучении любого процесса (физического, социального) приходится сталкиваться с необходимостью представлять его в качестве некоторой модели, т.е. в виде какого-то образа. Этот образ может быть заявлен в описательной форме (эпистолярный жанр), может изображаться в форме математического уравнения (формулы) или же показан как графическая картинка. Следовательно, сам оригинал (физический процесс, экономическое явление) заменяется некоторым аналогом, "эрзацем" (т.е. моделью). Такое создание "заместителя оригинала" и принято называть *аппроксимацией*.

Обычно под аппроксимацией (от лат.*approximatio* – приближение) понимают замену одного объекта другим, более известным и более простым, однако весьма близким к исходному по своему содержанию. В этом случае связь между исходным объектом (оригиналом) ***F*** и его приближенным представлением (моделью) ***f*** соответствует приближенному равенству $F \approx f$ (рис.3.1).

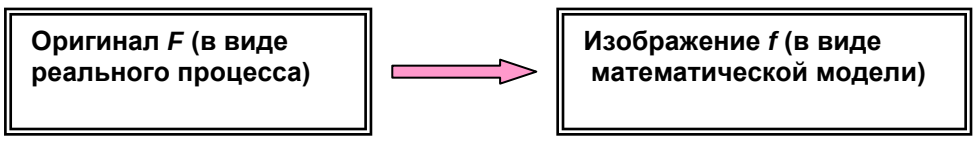


Рис.3.1. Схематическая связь между оригиналом и моделью объекта

Задача аппроксимации часто возникает при обработке результатов экспериментов, когда становится необходимым подобрать математическую мо-

дель изучаемого процесса, т.е. дать его аналитическое описание в виде так называемой *эмпирической формулы*.

При подборе эмпирических формул обычно используется *феноменологический* подход. Этот термин означает, что изучаемому процессу придается чисто *описательный вид*, при котором довольствуются только сведениями о *внешнем* характере этого процесса, но *игнорируется причинность* проявления рассматриваемой зависимости. В этом смысле феноменологический подход можно уподобить кибернетической модели "черного ящика". Как известно, при этом анализируется комбинация "вход-выход", т.е. характер влияния воздействующего фактора (аргумента) на исследуемый параметр (отклик или функцию). Однако содержимое "черного ящика" остается вещью в себе, т.е. физическая (или экономическая) природа процесса не обсуждается. Принципиальная особенность *физического* подхода состоит в том, что исследуемый процесс оценивается с позиций причин его проявления. Следовательно, если при феноменологическом подходе основной вопрос ставится в формулировке – "*Как произошло?*", то при физическом описании – "*Почему произошло?*" Тем самым феноменология дает чисто формальное, внешнее описание процесса, физический же подход основывается на выяснении его причин, его природы.

3.2. Выбор формул лучшего вида

При изучении связи показателей коммерческой деятельности применяются различного вида уравнения прямолинейной и криволинейной связи.

Формально могут возникать ситуации двух типов.

1. Вид функциональной зависимости *неизвестен*. В этом случае нужно решить предварительно задачу, направленную на отыскание подходящей функциональной зависимости. Это достаточно сложная задача, но она успешно решается современными средствами информационных технологий (программа Excel).

2. Вид функциональной зависимости *известен* и требуется только найти ее *параметры* (коэффициенты регрессии $\boldsymbol{b_0}$, $\boldsymbol{b_1}$, $\boldsymbol{b_2}$, …).

Термином *линейный регрессионный анализ* обозначают такое прогнозирование, которое описывается линейной взаимосвязью между исследуемыми переменными:

$$\boldsymbol{y=b_0+b_1x}.$$

В случае *криволинейных* зависимостей применяются математические функции следующего вида:

гиперболическая $y=b_0+b_1/x$;

показательная $y=b_0+b_1^x$;

степенная $y=b_0x^{b1}$;

параболическая $y=b_0+b_1x+b_2x^2$;

логарифмическая $y=b_0+b_1\lg x$;

экспоненциальная $y=b_0exp(b_1x)$ и другие.

Решение математических уравнений связи предполагает вычисление по исходным данным их параметров (*свободного члена* b_0 и *коэффициентов регрессии* $b_1, b_2, \ldots$).

При всем разнообразии эмпирических формул все же имеется вид аналитической зависимости, получивший широкое распространение. Им является уравнение регрессии в виде *многочленов (полинома)*, расположенных по восходящим степеням изучаемого фактора и одновременно линейных ко всем коэффициентам.

Такая формула имеет вид:

$$y=f(x)= b_0+b_1x+ b_2x^2+\ldots+b_m x^m ,$$

где $b_0, b_1, b_2, \ldots, b_m$ – коэффициенты, подлежащие определению.

Этот ряд – *сходящийся*, т.к. стремится к некоторому пределу.

Эмпирические формулы (аппроксимирующие уравнения) всегда имеют ограниченную область применения, которая не должна выходить за пределы имеющихся опытных данных.

Широкое применение аппроксимирующих уравнений объясняется следующими причинами.

1. Точное аналитическое выражение зависимости между исследуемыми величинами может оставаться неизвестным и поэтому по необходимости приходится ограничиваться приближенными формулами эмпирического характера.

2. Точная функциональная зависимость выражается формулой настолько сложной, что ее непосредственное применение при вычислениях было бы очень затруднительным.

Эмпирические формулы могут быть разнообразными, т.к. при выборе аналитической зависимости руководствуются не какими-то строгими теориями (физическими или экономическими), а ставят только одно условие – *возможно близкое соответствие значений, вычисленных по формуле, с опытными данными.* Таким образом, формально описание одного и того же процесса можно дать разными по виду уравнениями. Их пригодность оценивается только по одному критерию − наиболее точное предсказание экспериментального результата.

В эмпирическую формулу можно вводить различное число постоянных параметров (коэффициентов), величину которых нужно определить с большой точностью. Более удачными (удобными) следует считать уравнения с небольшим числом коэффициентов (не более 2−3). В противном случае возрастают трудности с применением таких формул.

3.3. Метод наименьших квадратов

Для определения коэффициентов уравнения регрессии $\boldsymbol{b}$ применяют разные методы (графический, метод средних), однако наибольшее распространение получил метод наименьших квадратов (МНК).

Пусть обсуждается некоторая зависимость $\boldsymbol{y=f(x)}$, которая отражает какой-то процесс, имеющий плавное течение и поэтому все параметры системы изменяются постепенно, без скачков. В этих случаях экспериментальные точки, нанесенные на графике, должны бы укладываться на некоторую плавную кривую (в частном случае, прямую). Однако на практике определенный разброс экспериментальных точек всегда наблюдается, что связано с изменчивостью (ошибками) регистрируемых измерений. Понятно, что такой разброс удалось бы избежать, если бы результаты измерений оказались совершенно свободными от ошибок, и тогда точки, отвечающие этим результатам, строго ложились бы на соответствующую плавную кривую (или прямую линию). Поэтому все процессы, которые имеют заведомо плавное течение, принято изображать также плавными кривыми, проводя их не через точки, а так, чтобы кривая проходила по возможности ближе ко всем точкам на графике.

Однако такое указание оставляет при построении кривых определенный произвол. Его частично можно устранить основным положением МНК:

сумма квадратов отклонений ε_i экспериментальных точек от кривой по вертикальному направлению, т.е. сумма квадратов величин ε_i, должна быть наименьшей ($\Sigma\varepsilon_i^2$ = минимум).

Или иначе – сумма квадратов отклонений известных (экспериментальных) значений исследуемой функции и соответствующих значений аппроксимирующей функции (теоретических показателей) должна быть наименьшей.

Довольно часто при описании аппроксимирующей функции ограничиваются простым видом полиноминальной зависимости, полагая ее линейной, т.е. в виде уравнения прямой $\boldsymbol{y=b_0+b_1x}$.

Здесь *свободный член* $\boldsymbol{b_0}$ характеризует *сдвиг* и равен тому значению $\boldsymbol{y}$, которое получается при $\boldsymbol{x}$=0, а *коэффициент* $\boldsymbol{b_1}$ определяет *наклон* линии.

Отыскание коэффициентов $\boldsymbol{b_0}$ и $\boldsymbol{b_1}$ осуществляется по *МНК.*

Пусть имеется $\boldsymbol{n}$ экспериментальных точек (*n* пар наблюдений): ($\boldsymbol{x_1}$, $\boldsymbol{y_1}$); ($\boldsymbol{x_2}$, $\boldsymbol{y_2}$);… ($\boldsymbol{x_n}$, $\boldsymbol{y_n}$). Введем следующие обозначения: $\boldsymbol{y_i}$ – это измеренные (экспериментальные) значения изучаемого параметра, а $\hat{\boldsymbol{y}}_\mathbf{i}$ – его теоретические (рассчитанные по уравнению) показатели.

Предположим, что экспериментальные точки на графике укладываются так, что по ним вполне возможно провести прямую линию (рис.3.2).

Значения функции $\hat{\boldsymbol{y}}_\mathbf{i}$ в этом случае можно записать в виде линейного уравнения:

$$\hat{\boldsymbol{y}}_\mathbf{i}= \boldsymbol{b_0} +\boldsymbol{b_1}\, \boldsymbol{x_i}\,,$$

Расстояние по ординате (вертикали) от точки $\boldsymbol{y_i}$ до прямой составит:

$$\boldsymbol{b_0} +\boldsymbol{b_1}\, \boldsymbol{x_i} - \boldsymbol{y_i} = \boldsymbol{\varepsilon_i},$$

где $\boldsymbol{b_0} +\boldsymbol{b_1}\, \boldsymbol{x_i} = \hat{\boldsymbol{y}}_\mathbf{i}$ – *рассчитанное (теоретическое)* значение функции; $\boldsymbol{y_i}$ – ее *измеренное (опытное)* значение и $\boldsymbol{\varepsilon_I}$ – *разница (расстояние)* между $\hat{\boldsymbol{y}}_\mathbf{i}$ и $\boldsymbol{y_i}$.

В соответствии с МНК полагаем, что искомая прямая будет наилучшей, если сумма квадратов всех расстояний $(\boldsymbol{b_0} +\boldsymbol{b_1}\, \boldsymbol{x_i} - \boldsymbol{y_i})^2=\boldsymbol{\varepsilon_i}^2$ окажется наименьшей.

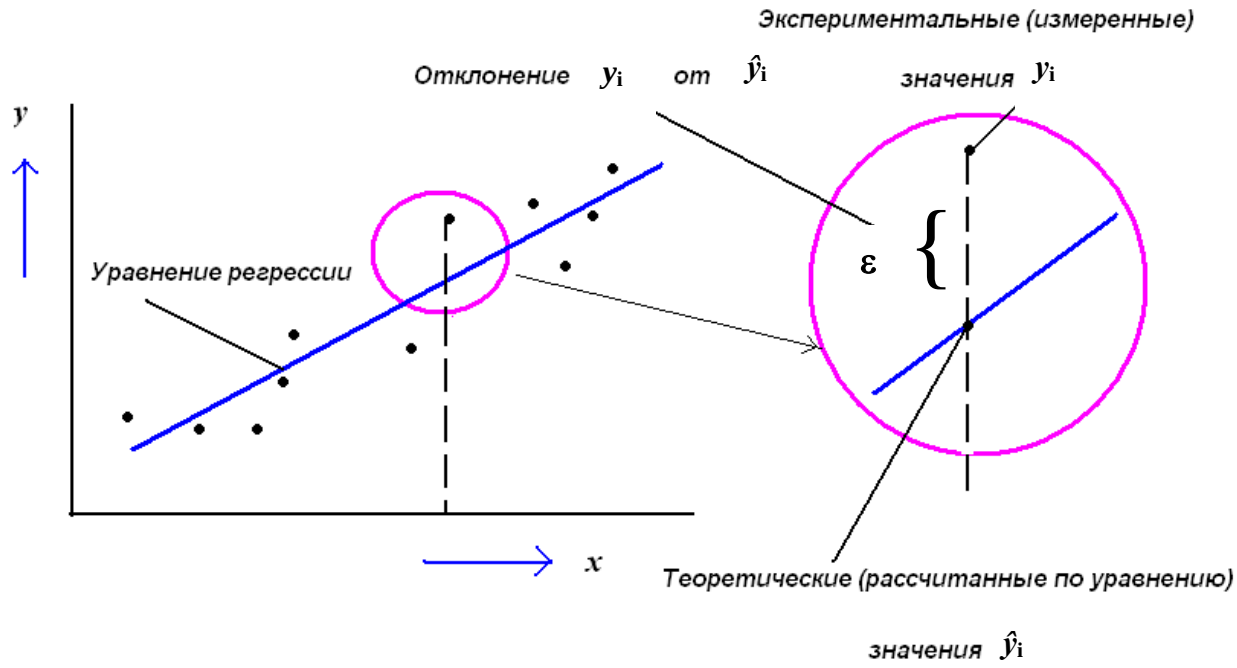


Рис. 3.2. Схематическое пояснение содержания метода наименьших квадратов

Минимум этой суммы ищется по правилам дифференциального исчисления. В результате для определения $\boldsymbol{b_0}$ и $\boldsymbol{b_1}$ используются следующие уравнения:

$$b_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}$$

Особенности МНК:

1. Этот метод *не* дает ответа на вопрос о том, какого вида функция лучше всего аппроксимирует конкретные экспериментальные точки.

Вид интересующей нас функции должен быть задан на основе каких-то физических или экономических соображений (либо специальным образом отыскан). МНК позволяет лишь выбрать, какая из прямых (парабол, экспо-

нент) является лучшей прямой (параболой, экспонентой) для прогнозирования.

2. Вычисления по МНК являются достаточно громоздкими, поэтому основная нагрузка – на компьютерные программы.

3. МНК является достаточно точным приемом и позволяет получить вполне надежные результаты. Одновременно он является *интерполяционным* методом, поскольку обеспечивает с определенной вероятностью предсказание любых значений y_i *в интервале* изученных значений x_i.

Напомним, что *экстраполяционный* метод (в отличие от интерполяционного) дает возможность предсказывать результаты *за пределами* изученной области.

После того, как уравнение регрессии найдено, необходимо определить его статистическую пригодность, т.е. выяснить, насколько оно верно (надежно) предсказывает в интервале $x_1; x_2; \ldots x_n$ экспериментальные результаты для y. Подобную оценку принято называть проверкой на значимость или адекватность.

3.4. Поиск уравнения регрессии

Рассмотрим на конкретном примере решение задачи по построению уравнения регрессии.

Студент Боб Деканкин решил в период летних каникул немного подзаработать, для чего устроился в контору "Ржавая подкова", занимающуюся сбором металлического лома от населения. Начальник конторы г-н Тютякин Фрол Макарович, преисполненный глубоким уважением к учености будущего дипломированного коммерсанта, попросил Боба проанализировать конкретные временные затраты на сбор (среди прочего металлолома) всяческих промышленных отходов и бытового старья из меди и её сплавов.

Боб Деканкин, знакомый с методом регрессионного анализа, решил взяться за порученное дело. В течение восьми рабочих дней он аккуратно регистрировал данные сбора медного металлолома. Картина получилась достаточно пестрой – были очень скудные по результатам дни, а какие-то оказывались весьма продуктивными. В целом это позволило представить в табличной форме (табл.3.1) основные итоги, указав для статистического массива $n=8$ следующие показатели: а) затраченное время (часы) и б) вес собранного металлолома (кг).

Т а б л и ц а 3.1

Результаты сбора медного лома в конторе "Ржавая подкова"

Время, затраченное на сбор медного лома, *x*, *ч*	**1,5 4,0 5,0 7,0 8,5 10,0 11,0 12,5**
Количество собранного металлолома *y*, *кг*	**5,0 4,5 7,0 6,5 9,5 9,0 11,0 9,0**

Итак, исследуется некоторая зависимость $y = f(x)$. Будем исходить из предположения, что эта зависимость описывается линейным уравнением. Об этом предварительно можно судить по виду построенного графика (рис.3.3).

3.4.1. Использование традиционных способов расчета

Вычисление на первом этапе проведем традиционным, а потому и самым утомительным способом, т.е. "вручную". Здесь нам в лучшем случае будет помогать лишь калькулятор.

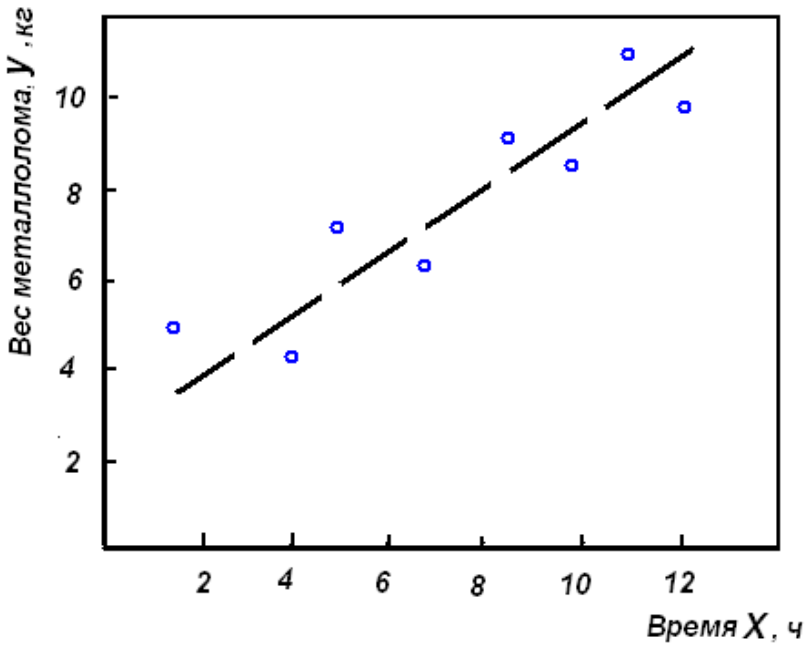


Рис.3.3. Графическое изображение исследуемой зависимости *y=f(x)*

Вычисление коэффициентов регрессии удобнее проводить в табличной форме. Для этого заполним табл.3.2, в которой, помимо исходных данных (их

мы расположим по столбцам), в графах 4-8 укажем вспомогательные расчетные данные.

Для проверки правильности вычисления в таблице можно использовать следующее выражение:

$$\Sigma(x+y)^2 = \Sigma x^2 + 2\Sigma xy + \Sigma y^2.$$

1. Определим среднее арифметическое для каждого ряда – для x и y. Они составят соответственно:

$$\bar{x} = 59{,}5/8 = 7{,}44 \text{ } ч \text{ и } \bar{y} = 61{,}5/8 = 7{,}69 \text{ } кг.$$

Значения полученных сумм подставляем в формулу для последующей проверки.

Получим:

2072,00=541,75+2·510,25+509,75;

2072,00=2072,00.

Следовательно, вычисления выполнены правильно.

Т а б л и ц а 3.2

Вспомогательная таблица для расчета коэффициентов регрессии

№ п/п	x	y	x^2	y^2	xy	$x+y$	$(x+y)^2$
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1	1,5	5,0	2,25	25,00	7,50	6,50	42,25
2	4,0	4,5	16,00	20,25	18,00	8,50	72,25
3	5,0	7,0	25,00	49,00	35,00	12,00	144,00
4	7,0	6,5	49,00	42,25	45,50	13,50	182,25
5	8,5	9,5	72,25	90,25	80,75	18,00	324,00
6	10,0	9,0	100,00	81,00	90,00	19,00	361,00
7	11,0	11,0	121,00	121,00	121,00	22,00	484,00
8	12,5	9,0	156,25	81,00	112,50	21,50	462,25
Итого	Σ=59,5	Σ=61,5	Σ=541,75	Σ=509,75	Σ=510,25	Σ=121,00	Σ=2072,00

2. Рассчитаем теперь коэффициенты b_0 и b_1 по известным формулам:

$$b_0 = \frac{\sum_{i=1}^{n} x_i^2 \sum_{i=1}^{n} y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} x_i y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2},$$

$$b_0 = \frac{541{,}75 \cdot 61{,}50 - 59{,}50 \cdot 510{,}25}{8 \cdot 541{,}75 - 59{,}50^2} = 3{,}73 \text{ кг.}$$

$$b_1 = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2},$$

$$b_1 = \frac{8 \cdot 510{,}25 - 59{,}50 \cdot 61{,}50}{8 \cdot 541{,}75 - 59{,}50^2} = 0{,}53 \text{ кг/ч.}$$

Таким образом, уравнение регрессии, т.е. формула, с некоторой вероятностью отображающая зависимость ***y*** от ***x***, имеет следующий вид:

$$\hat{y} = 3{,}73 + 0{,}53x.$$

3. Для проверки значимости (пригодности) полученного уравнения регрессии применяют специальные приемы. Такую проверку называют проверкой *адекватности модели*.

Для количественной проверки гипотезы об адекватности можно использовать так называемый ***F*** *–критерий* (*критерий Фишера*):

$$F = \frac{S_{ад}^2}{S_{общ}^2}.$$

Здесь $S_{ад}^2$ – *остаточная дисперсия* или *дисперсия адекватности*. Она характеризует величину *среднего разброса экспериментальных точек* Δy *относительно линии регрессии*, т.е. $\Delta y = y_i - \hat{y}_i$ (Δy есть ошибка в прогнозировании экспериментального результата на основании математической модели).

Остаточная дисперсия, таким образом, позволяет оценить *ошибку*, с которой *уравнение регрессии предсказывает фактический результат*. Следовательно, минимальная величина остаточной дисперсии должна свидетельствовать о более удачном выборе линии регрессии.

Вообще в статистике принято считать, что применение критерия минимальности остаточной дисперсии является вполне надежным способом отбора адекватных экономико-математических моделей.

Чтобы определить, велика или мала ошибка в предсказании эмпирических результатов, ее нужно сопоставить с некоторой *статистической величиной* (эталоном), принимаемой в качестве *критической*. Вот почему используется расчетный ***F-критерий***, который затем сравнивают с $\boldsymbol{F_{крит}}$.

Если $\boldsymbol{F_{расч} \le F_{крит}}$, то модель принимается *адекватной*, т.е. с заданной степенью достоверности (надежности) она верно предсказывает реальный результат. Если же $\boldsymbol{F_{расч} > F_{крит}}$, то вывод обратный – данное уравнение не может с заданной надежностью прогнозировать эмпирические данные.

Проверка адекватности модели по критерию Фишера дает возможность ответить на вопрос, *во сколько раз модель предсказывает результат хуже по сравнению с опытом.*

Остаточная дисперсия $S^2_{ад}$ рассчитывается путем деления остаточной суммы квадратов на число степеней свободы ***f*** по следующей формуле:

$$S^2_{ад} = \frac{\sum_{i=1}^{n} \Delta y^2}{f} .$$

Здесь число степеней свободы $\boldsymbol{f=n-(k+1)}$, где ***n*** – число опытов в эксперименте (т.е. может составлять объем случайной выборки); ***k*** – число изучаемых факторов.

Для однофакторного эксперимента имеем $\boldsymbol{f = n - 2}$ и тогда

$$S^2_{ад} = \frac{\sum_{i=1}^{n} \Delta y^2}{n-2} = \frac{\sum_{i=1}^{n} (\bar{y} - \hat{y})^2}{n-2} .$$

Вторая характеристика в формуле для расчета ***F***-критерия (знаменатель) – это так называемая *усредненная* или *общая дисперсия*. В качестве таковой принимается квадрат стандартной ошибки $S^2_{общ}$. Этот показатель фактически характеризует *случайную ошибку для всей выборки*, т.е. оценивает *несоответствие между конкретными* (текущими) *значениями* результата эксперимента и *средним арифметическим*.

Общая дисперсия рассчитывается как

$$S^2_{общ} = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{f} = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n-1}.$$

Вернемся к нашему примеру. Оценим статистическую пригодность полученного линейного уравнения. Показатель $S^2_{ад}$ удобно вычислять в табличной форме (табл.3.3).

Расчет проведем по формулам:

$$S^2_{ад} = \frac{\sum_{i=1}^{n}\Delta y^2}{n} = \frac{8{,}86}{8} = 1{,}11 \text{ и } S^2_{общ} = \frac{\sum_{i=1}^{n}(y_i - \hat{y})^2}{n} = \frac{35{,}05}{8} = 4{,}63.$$

Т а б л и ц а 3.3

Вспомогательная таблица для проверки уравнения на адекватность

№ п/п	x_i	y_i	$\hat{y}_i$=3,73+0,53x	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$	$y_i - \bar{y}_i$	$(y_i - \bar{y}_i)^2$
1	*2*	*3*	*4*	*5*	*6*	*7*	*8*
1	1,5	5,0	4,53	0,47	0,221	2,69	7,24
2	4,0	4,5	5,85	–1,35	1,822	3,19	10,18
3	5,0	7,0	6,36	0,62	0,384	0,69	0,48
4	7,0	6,5	7,44	–0,94	0,884	1,19	1,42
5	8,5	9,5	8,24	1,26	1,588	1,81	3,28
6	10,0	9,0	9,03	–0,03	0,001	1,31	1,72
7	11,0	11,0	9,53	1,44	2,074	3,31	10,96
8	12,5	9,0	10,35	–1,35	1,882	1,31	1,72
	Σ=59,5	Σ=61,5		Σ=0,12	Σ=8,86	Σ=15,51	Σ=37,05

Определим величину критерия Фишера: $F_{расч} = \frac{S^2_{ад}}{S^2_{общ}} = \frac{1{,}11}{4{,}63} = 0{,}24$.

Табличное значение определим для α=0,05, а также степеней свободы для числителя ($S^2_{ад}$) f_1 и знаменателя ($S^2_{общ}$) f_2. Они составят соответственно $f_1 = n - (k+1)$, где n – число опытов в эксперименте (объем случайной выбор-

ки); k – число изучаемых факторов x. Для однофакторного эксперимента имеем $f_1 = n - 2$.

Для второго показателя $f_2 = n - m$, где m – количество вычисленных констант для переменной y, которая соответствует среднему арифметическому $\bar{y}$ (т.е. $m=1$). Тогда $f_2 = n - 1$. В нашем случае имеем $f_1 = 8 - 2 = 6$ и $f_2 = 8 - 1 = 7$. В итоге для $\alpha = 0{,}05$ получим $F_{крит} = 3{,}87$ (*Приложение 5*).

Поскольку $0{,}24 < 3{,}87$, то с вероятностью 95% можно утверждать, что рассматриваемое уравнение адекватно, и оно способно с указанной достоверностью предсказывать экспериментальные результаты.

Если теперь возвратиться к самому обсуждаемому заданию, то можно заметить, что смышленый студент Боб Деканкин вполне управился с порученным делом. Он сообщил пытливому г-ну Тютякину, что на основании имеющихся опытных данных можно уверенно спрогнозировать (с надежностью 95%) результат сбора медного лома: так, например, за 8 часов работы это составит почти 8 кг ($3{,}73 + 0{,}53 \times 8 = 7{,}97$).

Пояснение. В литературе по статистики обычно используются два подхода к оценке $F_{расч}$: либо как отношение $S^2_{ад} / S^2_{общ}$, либо как $S^2_{общ} / S^2_{ад}$. Соответственно и статистический вывод на основании сравнения вычисленного F-критерия и эталонного $F_{крит}$ дается с учетом принятого соотношения. Нами рассматривается версия, когда $F_{расч} = S^2_{ад} / S^2_{общ}$; в то же время в компьютерной программе используется обратное отношение, т.е. $F_{расч} = S^2_{общ} / S^2_{ад}$. Это различие не носит принципиального характера. Важно только помнить, какой прием для анализа используется и, следовательно, каким образом дается надлежащее заключение.

3.4.2. Расчет с использованием компьютерной программы

А теперь покажем, что всю эту громоздкую и довольно затратную по времени процедуру можно весьма элегантным образом заменить услугами Excel.

Для этого на рабочем листе Excel предварительно организуем таблицу с исходными данными, в которой укажем содержимое табл.3.1. Причем саму таблицу построим по столбцам и поместим её в ячейках A1:C9. Итоговый результат показан на рис.3.4.

Далее будем действовать привычным образом:

1. В главном меню запустим серию команд ***Сервис/Анализ данных/Регрессия***.

2. В появившемся диалоговом окне заполним поля ввода данных для обоих параметров ***y*** и ***x***; для этого в каждое окно (***Входной интервал Y*** и ***Входной интервал X***) поместим наши данные, выделив их предварительно в соответствующих столбцах (напомним, что для функции ***y*** ее данные "сидят" в третьем столбце C2:C9, а для переменной ***x*** – во втором, т.е. B2:B9; при этом выделяются только те ячейки, которые содержат исключительно числовые показатели).

3. Отметим ***Уровень надежности*** (доверительную вероятность), равный 95%.

4. Укажем в окне вывода ***Выходной интервал*** ту ячейку, от которой будет формироваться весь блок получаемых статистических показателей, это D11; после чего – кнопка ***ОК***.

На рис.3.4 в собранном виде представлены все упомянутые элементы – исходная таблица (в верхнем левом углу), заполненное диалоговое окно ***Регрессия*** и, наконец, рассчитанные статистические показатели под заголовком "Вывод итогов".

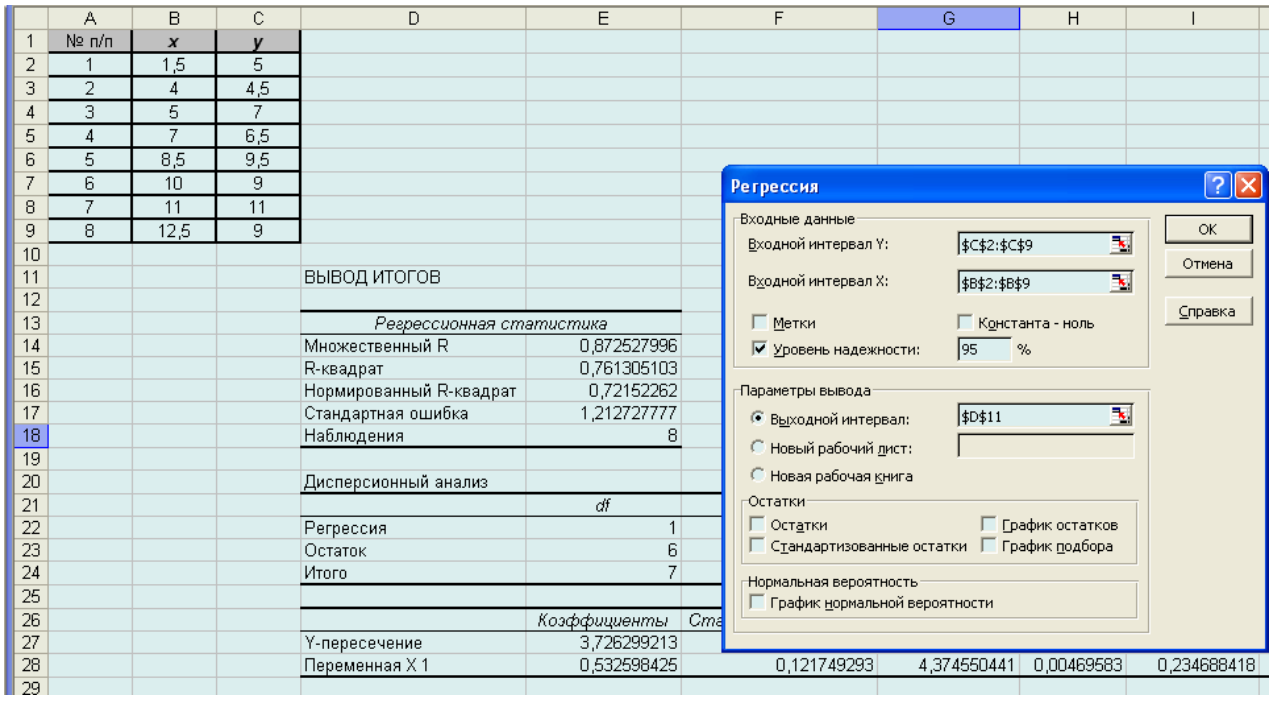


Рис. 3.4. Лист Excel с результатами расчета коэффициентов регрессии

Старательный Excel представил, как мы видим, разнообразные статистические материалы. Выберем, однако, из них пока только те, которые нам потребуются для заключительных рассуждений.

Интерес представляют показатели, которые именованы как "Коэффициенты". Один из них назван "Y-пересечение", а второй "Переменная X_1". Это и есть нужные нам коэффициенты регрессии – свободный член $\boldsymbol{b_0}$ и коэффициент $\boldsymbol{b_1}$ при аргументе $\boldsymbol{x}$. Если затем провести надлежащее округление до второго знака после запятой, то получим знакомые уже нам числа 3,73 и 0,53, которые были рассчитаны ранее, что называется, "на коленке".

Таким образом, на примере предложенной задачи мы познакомились с проведением регрессионного анализа различными приемами – весьма архаичным, требующим значительных и трудоемких расчетов, и компьютерным, легко и быстро позволяющим получить итоговый результат.

И последнее. После вычисления коэффициентов полученное уравнение регрессии надлежит подвергнуть проверке на адекватность. Такая процедура нами была выполнена, когда рассматривался первый вариант анализа. Однако и Excel позволяет сделать то же самое. Тот набор показателей, который мы проигнорировали, когда оценивали представленные данные под заголовком "Вывод итогов", как раз и призван сделать необходимые по этому поводу заключения. Ограничимся пока этими результатами (все же оценку пригодности уравнения мы дали, хотя и весьма обременительным способом), но более обстоятельно с этими возможностями Excel познакомимся в следующей главе.

3.5. Компьютерный подбор оптимального уравнения регрессии

Как отмечалось, анализируемый процесс может быть описан в математической форме, при этом используемые эмпирические формулы могут иметь различный вид. Поэтому выбор оптимального уравнения диктуется только одним соображением – данные теоретического расчёта (т.е. полученные из уравнения) должны в наибольшей степени совпадать с фактическими результатами.

Рассмотрим на конкретном примере возможность решения подобной задачи с использованием приложения Excel.

Обсуждается следующая задача:

Проведено N=8 опытов по изучению некоторой зависимости y=f(x) . В каждом варианте опыты повторялись ***n*** *раз, при этом число параллельных измерений для каждого конкретного варианта опыта могло заметно*

различаться (от 3 дублей до 5). Полученные экспериментальные данные представлены в табличной форме (табл.3.4).

Таблица 3.4

Результаты опыта по исследованию зависимости *y=f(x)*

Номер опыта *N*	Значение аргумента *x*	Значение функции *y* в повторных опытах				
		1	2	3	4	5
1	10	15	21	16	15	14
2	20	20	22	21	21	20
3	30	27	28	26	27	-
4	40	36	35	37	36	35
5	50	49	48	50	49	48
6	60	65	64	66	65	-
7	70	87	88	86	-	-
8	80	117	115	116	118	117

Надлежит выполнить следующие процедуры:

1. Провести первичную статистическую обработку экспериментальных данных с выявлением грубых промахов, определением среднеквадратичного отклонения и вычислением доверительного интервала для уровня значимости α=0,05.
2. Построить график рассматриваемой зависимости и подобрать для неё эмпирическую формулу.
3. Дать статистическую оценку подобранному уравнению.

Приступим к решению данного примера. Удобнеё всего придерживаться привычного алгоритма, т.е. будем указывать пошаговую последовательность наших манипуляций при работе с компьютером.

1. Сначала запустим Excel и откроем рабочий лист, в котором будет формироваться наш документ.

2. Теперь нужно ввести опытные данные. Для этого фактически придется повторить исходную таблицу, т. е. указать номера опытов, значения аргумента *x* и все значения функции *y* в параллельных опытах. Далеё добавим к нашей таблице ещё два столбца, в которые будут введены среднее арифметическое $\bar{x}$, среднеквадратичное отклонение S_n и

доверительный интервал Δx для каждого опыта, т.е. итоговые расчёты для каждой строки.

3. Приступим теперь к расчёту среднего арифметического и стандартного отклонения для каждой строки. Для этого нужно воспользоваться ***Мастером функций***. Перед запуском ***Мастера*** нужно выделить ту ячейку, в которую будет помещён искомый результат. Например, для определения среднего арифметического значения данных первой строки активизируем верхнюю ячейку предпоследней колонки. Затем запустим ***Мастер функций*** (кнопкой f_x или же в строке меню используем команды ***Вставка/Функция***).

Действия ***Мастера функций***:

– в появившемся диалоговом окне следует выбрать нужную функцию из списка (все функции разбиты на категории). Для этого в левой части панели (там перечислены категории) выберем требуемую под названием ***Статистические***, затем в правой части, где указаны функции, активизируем собственно нужную функцию ***Срзнач*** и далеё нажмём на кнопку ***ОК***;

– выделим теперь все ячейки первой строки, относящиеся к параметру *y*, т.е. это те ячейки, где расположены дубли первого опыта. После чего – кнопка ***ОК.*** Если теперь взглянуть на содержимое ячейки среднего арифметического, то там и будет указан полученный результат.

4. Далеё полагалось бы подобную процедуру проделать для всей матрицы (таблицы). Делаем следующеё. Выделим ячейку, где содержится среднее арифметическое, и протянем ***Маркер заполнения*** (маленький квадратик в правом нижнем углу) вдоль всей предпоследней колонки вниз. Что получим? Во всех соответствующих ячейках будут содержаться готовые расчётные данные среднего значения!

5. Подобные манипуляции проделываем и для следующей колонки – среднеквадратичного (стандартного) отклонения. Сделаем только одно пояснение. При работе с ***Мастером функций*** нужно будет активизировать функцию ***Стандотклон.***

Если окажется, что число знаков после запятой велико, то разрядность можно отрегулировать, активизировав соответствующую ячейку с данным числом, а затем в инструментальной строке использовать команду ***Уменьшить разряд.***

6. Для расчёта доверительного интервала используем те же опции посредством ***Мастера функций.*** Вся необходимая процедура становится

понятной из рис.3.5-3.6: нужно выделить функцию ***Доверит*** (рис.3.5), а затем в появившемся окне ***Аргументы функции*** заполнить запрашиваемые строки (рис.3.6). Для уровня значимости α укажем 0,05; затем введем значение уже рассчитанного стандартного отклонения S_n и число дублей n. Для первой строки это будет выглядеть так, как показано на рис.3.6.

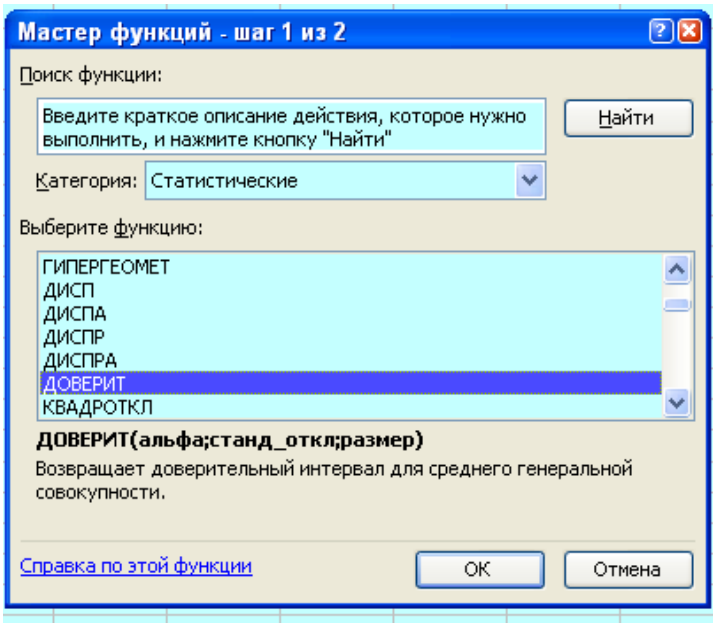


Рис.3.5. Поиск функции *Доверит*

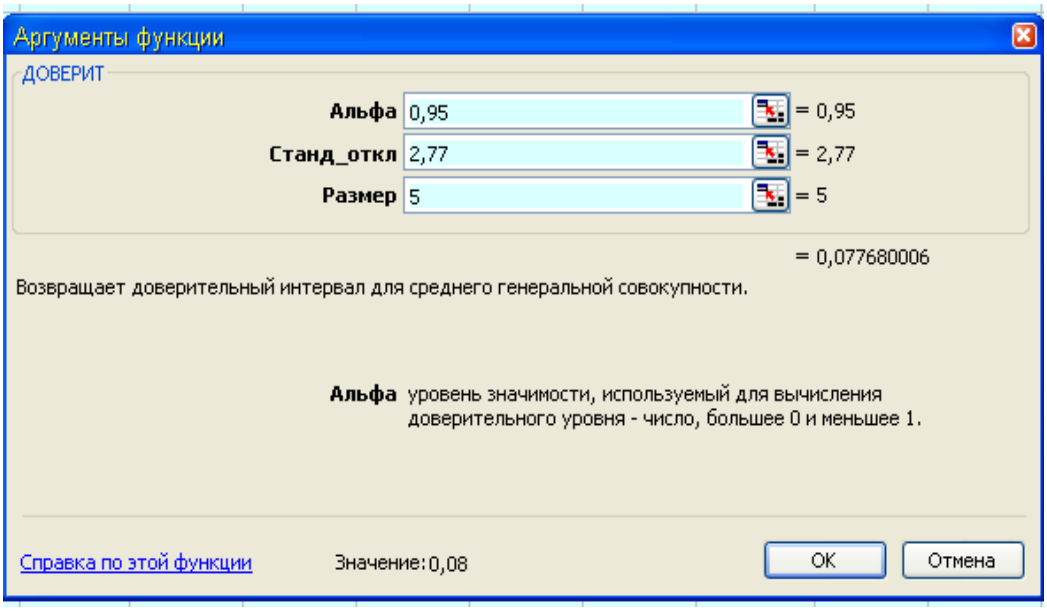


Рис.3.6. Панель для заполнения опции *Аргументы функции*

Тут следует обратить внимание на следующеё обстоятельство. При вычислении доверительного интервала нужно указывать число дублей, но их значения оказываются неодинаковыми – меняются от 3 (в 7-ом опыте) до 5 (в

5-и случаях). Поэтому такой расчёт нужно будет провести самостоятельно для каждой строки. Итоговый результат можно видеть на рис. 3.7.

O13

	A	B	C	D	E	F	G	H	I	J
1	Номер	Значение	Значение y в повторных опытах (дублях)							
2	опыта	x	1	2	3	4	5	$\bar{y}$	S_n	Δy
3	1	10	15	21	16	15	14	16,2	2,77	0,08
4	2	20	20	22	21	21	20	20,8	0,84	0,02
5	3	30	27	28	26	27		27	0,82	0,03
6	4	40	36	35		36	35	35,5	0,58	0,02
7	5	50	49	48	50	49	48	48,8	0,84	0,02
8	6	60	65	64	66	65		65	0,82	0,03
9	7	70	87	88	86			87	1,00	0,04
10	8	80	117	115	116	118	117	116,6	1,14	0,03

Рис.3.7. Экспериментальные данные после статистической обработки

7. Теперь пришел черёд проверить имеющиеся экспериментальные данные на наличие грубого промаха. Так, в первой серии настораживает результат 2-го измерения. Проверку надлежит провести по методу *максимального относительного отклонения*. Как делать – это уже знакомая процедура (см. раздел 1.11.2) Допустим, выполненные расчёты показали, что с вероятностью 95% этот результат следует признать грубым промахом (он не соответствует данной числовой совокупности). По этой причине его надлежит исключить из дальнейшего рассмотрения (т.е. в окончательном варианте число дублей первого опыта составит n=4).

8. Казалось бы, в очередной раз придется заняться расчётом среднего и стандартного отклонения (в данном случае для первой строки). Однако поступим следующим образом. Выделим ту ячейку, в которой содержится выскакивающий результат, и нажмём клавишу ***Delete***. Ячейка станет свободной, но при этом автоматически поменяются значения ***Срзнач*** и ***Стандотклон***.

Несколько иначе выглядит процедура определения доверительного интервала. Особенность структуры данной электронной таблицы такова, что изъятие выскакивающего результата не повлияет на изменение данных в ячейке. Причина та же – число дублей, как отмечалось, в разных опытах неодинаково. Поэтому для анализируемого варианта (1-й строки) придется отдельно вновь рассчитать ***Δx***.

Окончательный результат показан на рис.3.8. Незанятые (пустые) позиции в таблице означают отсутствие данных измерения в указанном повторном опыте или изъятие "нехорошего" (выскакивающего) результата.

Наконец, приступим к самому интересному этапу нашего задания – строим в графической форме анализируемую зависимость. В этом случае нам будет помогать ***Мастер диаграмм***. Он запускается либо нажатием клавиши на стандартной панели инструментов, либо через команды ***Вставка/Диаграмма*** в строке меню.

	A	B	C	D	E	F	G	H	I	J
1	Номер	Значение	Значение y в повторных опытах (дублях)							
2	опыта	x	1	2	3	4	5	$\bar{y}$	S_n	Δy
3	1	10	15		16	15	14	15	0,82	0,03
4	2	20	20	22	21	21	20	20,8	0,84	0,02
5	3	30	27	28	26	27		27	0,82	0,03
6	4	40	36	35		36	35	35,5	0,58	0,02
7	5	50	49	48	50	49	48	48,8	0,84	0,02
8	6	60	65	64	66	65		65	0,82	0,03
9	7	70	87	88	86			87	1,00	0,04
10	8	80	117	115	116	118	117	116,6	1,14	0,03

Рис.3.8. Итоговые данные

9.1. Запустим ***Мастер диаграмм*** и выполним рекомендации первого шага – выберем тип диаграммы. В появившемся окне, в левой его части, высветим тип диаграммы – ***График.*** Здесь же, нажав кнопку Просмотр результата, можно будет посмотреть, как станут выглядеть наши данные на диаграмме выбранного типа.

9.2. Нажмём на клавишу ***Далеё*** и перейдём, следовательно, ко второму шагу. В окне будет активизирована вкладка ***Диапазон данных***. Теперь в кнопке ***Ряды в*** следует указать, что наши данные представлены в ***Столбцах.*** Отметим, что на оси ординат будут указаны заданные численные значения аргумента, а вот на оси абсцисс пока содержатся некие нейтральные показатели типа 1, 2, 3 и проч.

9.3. В пределах окна второго шага высветим вкладку ***Ряд*** и в строке ***Подписи оси X*** ставим маркер. После чего сдвинем это кно так, чтобы можно было увидеть ту колонку таблицы, где ″сидят″ наши данные по аргументу x. Выделим весь этот столбец – на графике по оси абсцисс появятся фактические значения аргумента.

9.4. Совершим затем следующий, третий шаг (клавиша ***Далеё***). Он позволяет указать конкретные параметры диаграммы. Запустив вкладку ***Заголовки***, присвоим название диаграмме ("Экспериментальная зави-

симость"), а также отметим оси координат (записываем символы X и Y). По желанию можно "уокрасить" график – добавить или убрать сетку (вкладка ***Линии сетки***), дать необходимые комментарии к графику (вкладка ***Легенда***).

9.5. Последний шаг – укажем, где желательно разместить график. Для этого вновь нажимаем на кнопку ***Далеё*** и отмечаем место расположения его – на имеющемся листе или же отдельном. После завершения этой процедуры последняя приятная операция – прикоснуться к кнопке ***Готово***. Получаем график, имеющий вид, представленный на рис.3.9.

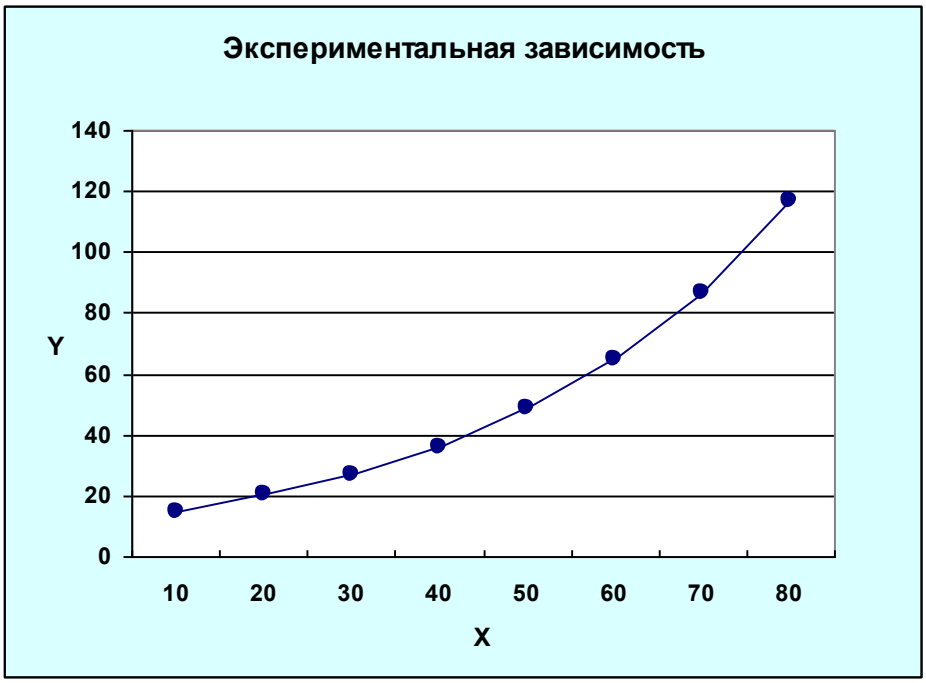


Рис.3.9. Графика исследуемой зависимости

Может оказаться, что габариты графика нас решительно не устраивают. Для придания ему более благообразного и удобного вида выделим ***Область диаграммы*** (должны появиться по периметру маркеры-засечки) и поменяем размеры (указатель мыши подведём к маркерам – должны возникнуть двойные стрелки, которые и нужно перемещать). Схожим образом можно изменить габариты самого графика (в пределах имеющейся области диаграммы), выделив ***Область построения диаграммы***

В случае надобности можно также исправить вид осей координат, изменив шрифт или размер цифр шкалы, добавив промежуточные засечки. Для этого нужно подвести стрелку мыши к выбранной оси и щёлкнуть правой клавишей. Появится окно Формат оси, которое после его активизации и позволит осуществить нужные манипуляции.

10. Заключительная процедура нашей работы (своеобразный "высший пилотаж" статистической обработки результатов измерения) – это аналитическое описание построенной экспериментальной зависимости. Для этого подведём стрелку мыши к линии графика и щёлкнем правой клавишей. Появится окно ***Формат рядов данных.*** Выделим опцию ***Добавить линию тренда***, в результате появится всплывающеё окно ***Линия тренда***. На вкладке ***Тип*** выберем похожий на нашу кривую график-шаблон. Для данного случая вполне подходящей оказывается полиноминальную зависимость второй степени (квадратное уравнение). Перейдём затем к вкладке ***Параметры*** и укажем засечками команды ***Показать уравнение на диаграмме*** и ***Поместить на диаграмме величину достоверной аппроксимации*** R^2. После нажатия клавиши ***ОК*** график примет окончательный вид маленького компьютерного шедевра (рис.3.10). Отметим, что наша экспериментальная кривая практически полностью совпала с теоретической. Это и неудивительно, поскольку аппроксимирующий коэффициент близок к 1 – идеальное соответствие!

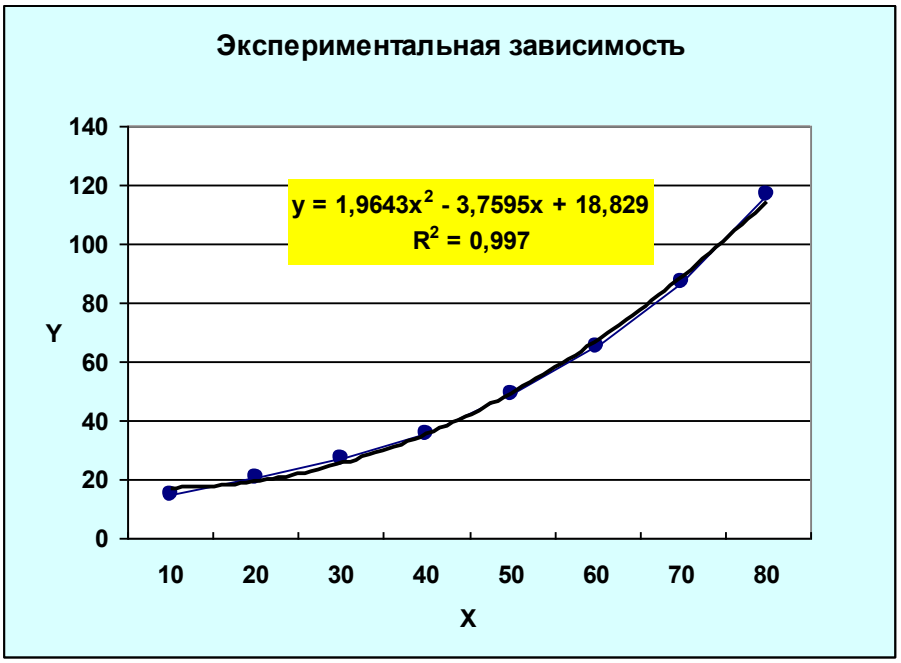


Рис.3.10. Окончательный вид аналитической зависимости

Фактически данную работу на этом можно считать и законченной. Однако сделаем ещё некоторые оценки. Дело в том, что мы, пользуясь эталонным набором кривых аналитических зависимостей (вкладка ***Тип*** из окна ***Формат рядов данных***), удачно выбрали полиноминальный вид функции.

Количественно об этом можно судить по величине аппроксимирующего коэффициента R^2. Можно вполне обоснованно показать, что выбранная зависимость является, похоже, наилучшей. С этой целью для наглядности проверим и другие функции, нанеся на график соответствующую линию тренда, а также показав получаемые уравнения регрессии и величины коэффициента R^2.

Такую процедуру нетрудно выполнить, после чего для рассмотренного примера полученные показатели R^2 для разных уравнений регрессии будут иметь следующий вид:

экспоненциальная – R^2 =0,999;

полиномиальная – R^2 =0,997;

линейная – R^2 =0,925;

степенная – R^2 =0,922;

логарифмическая – R^2 =0,730.

Как видно, обсуждаемая зависимость $y=f(x)$ лучше всего, как и предполагалось, описывается экспоненциальным уравнением. Этот вывод базируется не только на визуальных впечатлениях (вполне адекватное совпадение экспериментальной кривой и линии тренда), но и на строгом количественном расчёте с использованием статистического коэффициента R^2. Вместе с тем можно утверждать, что ещё более обоснованным представляется описание аппроксимации в виде экспоненциального уравнения, поскольку в этом случае рассчитанное значение коэффициента фактически оказывается равным единице.

4. МНОЖЕСТВЕННАЯ РЕГРЕССИЯ

Сложные проблемы всегда имеют простые, легкие для понимания неправильные решения.

(Закон Мэрфи)

До сих пор нами рассматривалась ситуация, когда на зависимую переменную (функцию) воздействовал только *один* фактор (аргумент). Подобное прогнозирование принято называть *простой регрессией*. Такие зависимости мы уже рассмотрели ранее.

Однако в подавляющем большинстве случаев приходится иметь дело с экспериментальными данными, касающимися влияния *более чем одного* фактора. Прогнозирование *единственной* переменной y на основании *нескольких* переменных x_i называется *множественной регрессией*. В этом случае математическая модель процесса представляется в виде уравнения регрессии с несколькими переменными величинами, т.е. $y = f(b_0, \ldots, x_k)$.

Общий вид уравнения множественной регрессии обычно стараются представить в форме линейной зависимости:

$$y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k,$$

где b_0 – свободный член (или сдвиг); $b_1, b_2, \ldots, b_k$ – коэффициенты регрессии, которые подлежат вычислению методом наименьших квадратов.

При анализе уравнения множественной регрессии (как и в случае простой регрессии) также используется такое понятие как *ошибка прогнозирования* Δy. Под этим понимается разность между *рассчитанным (теоретическим)* значением функции $\hat{y}_i$ и ее *измеренным (опытным)* значением y_i, т.е. $\Delta y = \hat{y}_i - y_i$.

Статистический вывод о пригодности (значимости) уравнения обычно проверяется в следующей последовательности.

1. Сначала проводится *общая* проверка методом *F-теста*, целью которой является выяснение, объясняют ли x-переменные значимую долю вариации y, т.е. превалирует ли влияние факторов x_i на изменение функции y над её колебаниями случайного порядка; если регрессия *не* является значимой, то говорить больше не о чем.

2. Если регрессия оказывается значимой, то можно продолжить анализ, используя *t-тесты* для *отдельных* коэффициентов регрессии; в этом случае

пытаются выяснить, насколько значимой является влияние той или иной переменной x на параметр y *при условии, что все другие факторы* x_i *остаются неизменными*. Построение доверительных интервалов и проверки гипотез на адекватность для отдельного коэффициента регрессии основывается на определении стандартной ошибки. Каждый коэффициент регрессии имеет свою стандартную ошибку S_{b1}, S_{b2},…, S_{bk}.

Рассмотрим конкретный пример.

Замечательная корова кота Матроскина радовала превосходными надоями, и поэтому он вознамерился излишки молока продавать. При этом Матроскин решил выяснить, каким образом объём ежедневной продажи молока y *(литров в день) зависит от а) присутствия среди покупателей бабушек с внучками (их доля от общего числа покупателей* x_1*, %) и б) участия в коммерции пса Шарика (относительное время* x_2*, когда он помогал работать за прилавком, %). Тщательные наблюдения Матроскин вел в течение 20 рабочих дней, результаты которых представил в табличной форме (табл.4.1). При этом порядковые номера торговых дней были расположены в случайном порядке и никак формально не отражали какое-либо внятное изменение объема продажи молока.*

Требуется помочь коту Матроскину:

– написать уравнение множественной регрессии;

– оценить статистическую значимость уравнения;

– определить значимость коэффициентов регрессии и пояснить характер влияния исследуемых факторов.

Если поставленную задачу сформулировать в более понятных для кота категориях, то нужно выяснить, влияют ли указанные факторы на его коммерческую деятельность в области молочного бизнеса, а если это так, то насколько ощутимо.

Т а б л и ц а 4.1

Исходные данные об эффективности продажи молока

Порядковый номер дня продажи	*y*, *л/день*	x_1, %	x_2,%	Порядковый номер дня продажи	*y*, *л/день*	x_1, %	x_2,%
1	6	40	30	11	7,5	50	35
2	4,6	20	33	12	7,7	37	30
3	4,4	31	20	13	7,3	50	40
4	4,5	32	25	14	7	38	42
5	5,5	34	29	15	6,7	50	39
6	4,8	35	20	16	5,7	35	35
7	5,1	37	21	17	6	46	36
8	5,2	32	20	18	6,4	49	38

9 10	7 5,3	39 35	35 30	19 20	7,1 6,3	51 45	41 34

4.1. Расчет коэффициентов регрессии и представление уравнения множественной регрессии

Итак, нам надлежит выполнить предложенную задачу. Вся прелесть исходной ситуации состоит в том, что по представленным данным решительно невозможно обнаружить какую-то сколь-нибудь заметную тенденцию. Поэтому решение задачи постараемся обеспечить с использованием компьютерных программ в режиме Windows.

Запускаем Excel и затем воспроизводим в табличной форме имеющиеся исходные результаты (табл.4.1). В данном случае все экспериментальные данные (по каждой позиции) представляем в виде самостоятельных колонок. Размещаем всю таблицу в ячейках от A1 до D21, при этом сами исходные данные (т.е. для y и x_1, x_2) будут находиться в диапазоне B1: D21.

	A	B	C	D
1	Номер	Y	X1	X2
2	1	6	40	30
3	2	4,6	20	33
4	3	4,4	31	20
5	4	4,5	32	25
6	5	5,5	34	29
7	6	4,8	35	20
8	7	5,1	37	21
9	8	5,2	32	20
10	9	7	39	35
11	10	5,3	35	30
12	11	7,5	50	35
13	12	7,7	37	30
14	13	7,3	50	40
15	14	7	38	42
16	15	6,7	50	39
17	16	5,7	35	35
18	17	6	46	36
19	18	6,4	49	38
20	19	7,1	51	41
21	20	6,3	45	34

Рис.4.1. Лист Excel с исходными табличными результатами

После этого получим сводную таблицу основных статистических характеристик для функции y. Для этого воспользуемся известным методом анализа данных – программой ***Описательная статистика***.

Предпримем следующие шаги.

1. В главном меню выберем последовательно пункты ***Сервис/Анализ данных/Описательная статистика***, после чего щелкнем по кнопке ***ОК***.

2. Заполним диалоговое окно для ввода данных и параметров вывода. Для этого проделаем следующие манипуляции (рис.4.2):

– укажем ***Входной интервал*** (в виде абсолютных ссылок $B$1:$D$21), т.е. адресуем все ячейки, в которых находятся значения функции ***y*** и аргументов x_1, x_2 ;

– отметим способ ***Группирования*** (в нашем случае по столбцам);

– установим флажок для ***Метки***, показывающий, что первая строка содержит название столбца;

– выделим ***Выходной интервал***, для этого достаточно указать левую верхнюю ячейку будущего диапазона ($F$1);

– установим флажки, показывающие, что нам нужна информация в виде ***Итоговой статистики***, а также ***Уровень надежности***, равный 95%; после чего – кнопка ***ОК***.

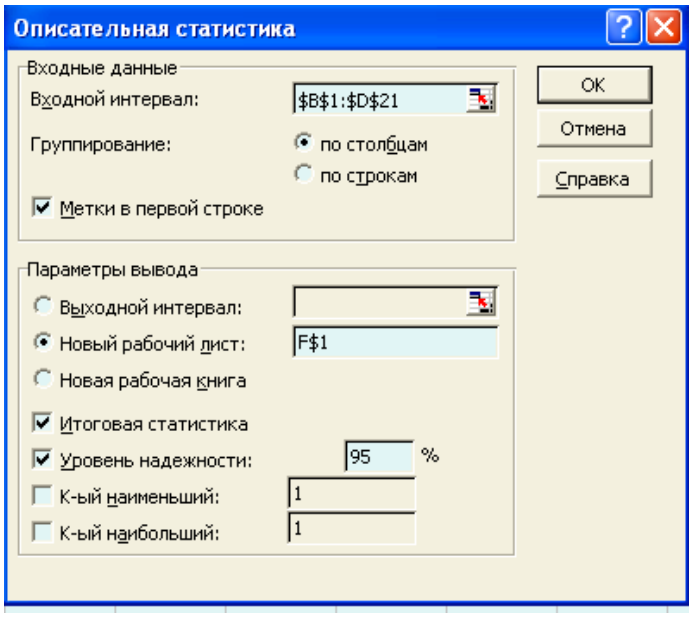


Рис.4.2. Диалоговое окно ввода параметров *Описательная статистика*

Полученные результаты статистического расчета показаны на рис.4.3 в виде соответствующего листа Excel.

	E	F	G	H	I	J	K
1		*Y*		*X1*		*X2*	
2							
3		Среднее	6,01	Среднее	39,30	Среднее	31,65
4		Стандартная ошибка	0,24	Стандартная ошибка	1,85	Стандартная ошибка	1,62
5		Медиана	6,00	Медиана	37,50	Медиана	33,50
6		Мода	6,00	Мода	35,00	Мода	30,00
7		Стандартное отклонение	1,06	Стандартное отклонение	8,26	Стандартное отклонение	7,25
8		Дисперсия выборки	1,12	Дисперсия выборки	68,22	Дисперсия выборки	52,56
9		Эксцесс	-1,30	Эксцесс	-0,11	Эксцесс	-0,94
10		Асимметричность	0,01	Асимметричность	-0,24	Асимметричность	-0,45
11		Интервал	3,30	Интервал	31,00	Интервал	22,00
12		Минимум	4,40	Минимум	20,00	Минимум	20,00
13		Максимум	7,70	Максимум	51,00	Максимум	42,00
14		Сумма	120,10	Сумма	786,00	Сумма	633,00
15		Счет	20,00	Счет	20,00	Счет	20,00
16		Уровень надежности(95,0%)	0,4952	Уровень надежности(95,0%)	3,87	Уровень надежности(95,0%)	3,39
17							

Рис.4.3. Лист Excel с результатами расчета статистических показателей

Из представленного набора статистических показателей выберем те, которые нам потребуются для последующего анализа – среднее арифметическое и стандартное отклонение (среднеквадратичное отклонение) S_n.

В табл.4.2 приведены указанные статистические показатели для функции y и обеих переменных x_1 и x_2. Отметим, что для функции y её среднее арифметическое $\bar{y}$ составляет 6,01, а и стандартное отклонение S_n равно 1,06.

Т а б л и ц а 4.2

Статистические показатели для функции y и переменных x_1 и x_2

Показатели	**Y**	**X_1**	**X_2**
Среднее арифметическое	**6,01**	**39,3**	**31,65**
Стандартное отклонение S_n	**1,06**	**8,26**	**7,25**

3. Расчет показателей регрессии также исполняется по компьютерной программе. Для ее запуска исполним следующие команды:

– в главном меню выберем пункты ***Сервис/Анализ данных/Регрессия***, после чего щелкнем по кнопке ***ОК***;

– заполним диалоговое окно ввода данных для параметра y и обеих характеристик x_1 и x_2; для этого в каждое окно (***Интервал Y*** и ***Интервал X***) поместим наши данные, выделив их предварительно в соответствующих столбцах (напомним, что для функции y её данные "сидят" во втором столбце B2:B21, а для переменных x_1 и x_2 – в третьем и четвертом, т.е. C2:D21; заметим, что при этом выделяются только те ячейки, которые содержат исключительно числовые показатели);

– выделим в текстовом поле ***Выходной интервал*** ту ячейку, от которой будет формироваться весь блок получаемых статистических показателей; при этом укажем другой лист – ***Лист 2***;

– после чего кнопка ***ОК***.

Заполненное диалоговое окно для программы ***Регрессия*** представлено на рис.4.4.

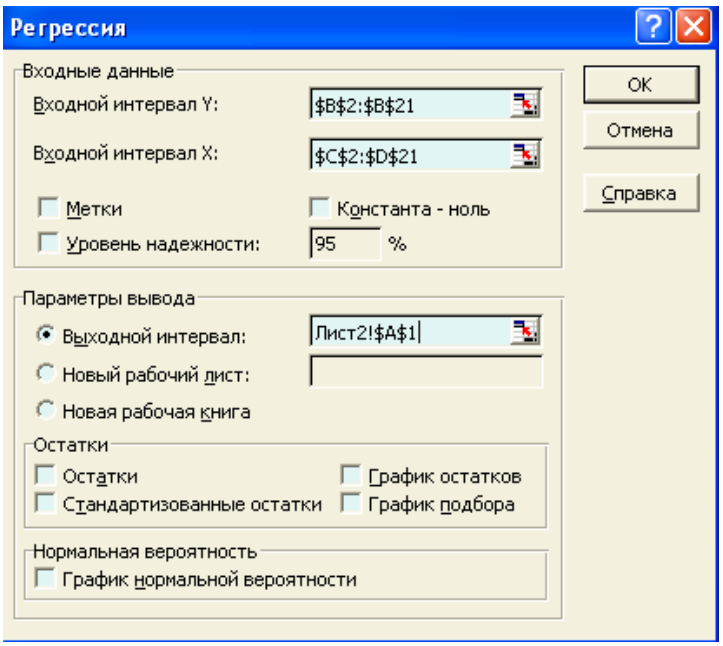


Рис.4.4. Диалоговое окно ввода параметров *Регрессия*

Как видно, мы получили набор разнообразных статистических материалов (рис.4.5). Выберем, однако, из них такие, которые нам потребуются для последующего анализа.

	A	B	C	D	E	F	G	H
1	ВЫВОД ИТОГОВ							
2								
3	*Регрессионная статистика*							
4	Множественный R	0,82						
5	R-квадрат	0,67						
6	Нормированный R-квадрат	0,63						
7	Стандартная ошибка	0,65						
8	Наблюдения	20,00						
9								
10	Дисперсионный анализ							
11		*df*	*SS*	*MS*	*F*	*Значимость F*		
12	Регрессия	2	14,18	7,09	16,99	8,84E-05		
13	Остаток	17	7,09	0,42				
14	Итого	19	21,27					
15								
16		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-Значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*
17	Y-пересечение	1,61	0,77	2,09	0,05	-0,02	3,23	-0,02
18	Переменная X 1	0,06	0,02	2,59	0,02	0,01	0,11	0,01
19	Переменная X 2	0,07	0,03	2,57	0,02	0,01	0,12	0,01
20								

Рис.4.5. Лист Excel с результатами расчета статистических показателей регрессии

Для этого организуем табл.4.3, в которой поместим *расчетные* значения коэффициентов регрессии, стандартную ошибку, величины ***t***-критерия и показатели уровня значимости α. Укажем также (ниже таблицы) рассчитанные показатели для самой функции ***y***.

Т а б л и ц а 4.3

Данные регрессионной статистики

Независимая переменная	Коэффициент	Стандартная ошибка	*t*	*p* (или α)
Свободный член	1,61	0,77	2,09	0,05
X_1	0,06	0,23	2,59	0,02
X_2	0,07	0,03	2,57	0,02

Для функции ***y***: $S_{\overline{Y}}$ = 0,65; ***R***-квадрат = 0,67; ***R***-квадрат (нормированный) = 0,63

Таким образом, для рассматриваемого примера уравнение регрессии (или уравнение прогнозирования) будет иметь следующий вид:

$\hat{y}$ (*объем продажи молока, л/день*) $= b_0 + b_1x_1 + b_2x_2 =$
=1,61 + 0,06 (*доля среди покупателей бабушек с внучками, %*) +
+ 0,07 (*относительный вклад участия в торговле Шарика, %*).

Запишем полученное уравнение в окончательной редакции:

$$\hat{y} = 1{,}61+0{,}06\, x_1+0{,}07\, x_2.$$

Теперь займемся статистическим анализом этого уравнения регрессии.

4.2. Интерпретация коэффициентов регрессии

Свободный член (сдвиг) b_0, равный 1,61, формально надлежит понимать следующим образом: объём продажи молока котом Матроскиным, когда отсутствуют среди покупателей бабушки с внучками и нет компаньона Шарика (занят фотоохотой), составляет 1,61 литров в день. Однако мы полагаем, что в указанной совокупности исходных данных нет подобных примеров (всегда среди покупателей окажутся бабушки с внучками, а Шарик помо-

гает ежедневно). Поэтому сдвиг b_0 следует обсуждать как вспомогательную величину, необходимую для получения оптимальных прогнозов, и не истолковывать её столь буквально.

Коэффициенты регрессии b_1 и b_2 следует рассматривать как степень влияния каждой из переменных (присутствие бабушек с внучками и вклад коммерческого таланта Шарика) на размер продажи, если все другие независимые переменные остаются неизменными. Так, коэффициент b_1, равный 0,06, указывает, что (при прочих равных условиях) повышение доли бабушек с внучками на 1% приводит к возрастанию продажи молока на 0,06 литров в день. Относительно коэффициента b_2 можно заметить, что увеличение относительного участия Шарика на 1% приводит также к повышению продажи и этот прирост составляет почти такую же величину – 0,07 л/день.

Ещё раз заметим, что все названные коэффициенты регрессии отражают влияние на исследуемый параметр y только какой-то одной переменной x при непременном условии, что все другие переменные (факторы) не меняются. Скажем, применительно к коэффициенту b_2 это нужно понимать так: указанное влияние коммерческой помощи Шарика проявляется при условии, когда сохраняется среди покупателей неизменной доля старушек с внучками.

4.3. Ошибки прогнозирования (определение качества регрессионного анализа)

Можно воспользоваться двумя приемами для оценки добротности выполненного нами регрессионного анализа. В статистике для этого используют:

– *стандартную ошибку* (S_y), которая дает представление о приблизительной величине ошибки прогнозирования;

– *коэффициент детерминации* (R^2), указывающий, какой процент вариации функции y объясняется воздействием факторов x_i.

Рассмотрим оба подхода более подробно.

1. Результаты статистического расчета показывают, что стандартная ошибка для функции составляет 0,65. Этот результат применительно к нашему примеру следует рассматривать следующим образом: фактическая величина объёма продаж молока отличается от прогнозируемых показателей не более чем на 0,65 л/день. Однако ценность этого показателя невелика, если не назвать, какова же надежность этого утверждения. При условии сохранения нормального распределения можно полагать, что примерно 2/3 фактиче-

ских данных будут находиться в пределах $S_{\bar{y}}$ от прогнозируемых показателей; примерно 95% – в пределах $2S_{\bar{y}}$ и т.д.

Эта стандартная ошибка $S_{\bar{y}}$, равная 0,65, указывает отклонение фактических данных от прогнозируемых на основании использования воздействующих факторов x_1 и x_2 (влияние среди покупателей бабушек с внучками и высокопрофессионального вклада Шарика). В то же время мы располагаем обычным стандартным отклонением S_n, равным 1,06 (см. табл.4.2), которое было рассчитано для одной переменной, а именно: сами текущие значения y_i и величина среднего арифметического $\bar{y}$ (оно равно 6,01). Легко видеть, что $S_{\bar{y}} < S_n$; следовательно, ошибки прогнозирования, как правило, оказываются меньшими, если использовать уравнение регрессии (т.е. учитывается вклад факторов x_1 и x_2), а не ограничиваться только значением $\bar{y}$.

Сказанное можно истолковать следующим образом. Если бы нам ничего не было известно про переменные x_1 и x_2, то в качестве оптимальной приблизительной величины среднего уровня продаж пришлось бы использовать показатель $\bar{y}$ = 6,01 л/день и полагать, что наши прогнозы дают ошибку S_n, равную 1,06 л/день. Однако если нам известны такие характеристики, как влияние особой категории покупателей (бабушки с внучками) и роль высококвалифицированного Шарика, то для прогнозирования можно воспользоваться уравнением регрессии. В этом случае наши предсказания будут давать ошибку уже примерно в 0,65 л/день.

Такое сокращение погрешности прогнозирования с 1,06 до 0,65 и является одним из преимуществ использования регрессионного анализа.

2. Если вновь обратиться к нашему примеру, то коэффициент детерминации R^2 (на рис.4.5 Excel его подает как R-квадрат) равен 0,67 или же составляет 67%. Этот результат следует толковать так: все исследуемые воздействующие факторы (влияние особой категории покупателей и коммерческого таланта Шарика) объясняют 67% вариации анализируемой функции (объема проданного молока). Остальное же (33%, что весьма прилично!) остается необъясненным и может быть связано с влиянием других, неучтенных факторов.

Для нашего примера показатель R^2 (67%) считается умеренным и поэтому можно полагать, что именно эти два фактора в данном конкретном случае оказываются достаточно влияющими.

4.4. Проверка значимости модели

Итак, нами получено уравнение множественной регрессии, коэффициенты которого b_i формально показывают, как и в каком направлении действуют (пока лишь вероятно!) исследуемые факторы x_i, и какой процент изменчивости функции y объясняется влиянием именно этих факторов.

Теперь нам надлежит определить статистическую значимость полученного аналитического выражения. Принято придерживаться следующей последовательности:

1. Сначала выполняется общая проверка полученного уравнения на пригодность.
2. Если результат оказался положительным (уравнение значимо), то проверяют на значимость уже каждый коэффициент уравнения регрессии b_i.
3. Дается сравнительная оценка степени влияния каждого из анализируемых факторов x_i.

4.4.1. Проверка на адекватность уравнения регрессии

Статистическую оценку полученного уравнения (так называемый *статистический вывод*) принято начинать с проведения F-теста, целью которого является выяснение способности исследуемых факторов x_i объяснить значимую часть колебания функции y. Этот тест используется как своеобразные "входные ворота" в статистический вывод: если результат теста значим, то связь, следовательно, существует и можно приступать к её исследованию и объяснению. Если показатель теста незначим, то заключение лишь одно – мы имеем дело с набором случайных чисел, никак не связанных между собой. И больше делать нечего, так как нет предмета для анализа…

Заметим при этом, что сам формальный факт отсутствия значимости на деле может и не соответствовать отсутствию взаимосвязи как таковой. Просто в указанных обстоятельствах у нас не хватило экспериментальных данных доказать, что такая связь вообще-то есть. Иначе говоря, она может и быть, но из-за малого размера выборки или случайности нам не удалось её доказать на основании тех опытных данных, которые фактически были в нашем распоряжении.

Использование так называемой *нулевой гипотезы* для F-теста означает, что между переменными x_i и y значимая связь *отсутствует*. Следовательно, признается, что параметр y является чисто случайной величиной и значения переменных x_i не оказывают на него никакого систематического влияния. Применительно к уравнению регрессии это утверждение можно трактовать как случай, когда *все* коэффициенты уравнения равны *нулю*.

В свою очередь *альтернативная гипотеза* ***F***-теста говорит о том, что между параметром ***y*** и переменными x_i существует определенная прогнозирующая взаимосвязь. Следовательно, параметр ***y*** уже не является чисто случайной величиной и должен зависеть хотя бы от одной из переменных x_i. Тем самым альтернативная гипотеза настаивает на том, что, по крайней мере, один из коэффициентов регрессии отличен от нуля. Как видно, здесь принимается во внимание следующее обстоятельство: совершенно необязательно, чтобы каждая ***x***-переменная влияла на параметр ***y***, вполне достаточно, чтобы влияла хотя бы одна из них.

Для выполнения ***F***-теста воспользуемся результатами компьютерного расчета, который исполнил замечательный Excel. Здесь обычно рекомендуются следующие приемы.

1. *Решение принимается на основе критерия Фишера.*

Это достаточно традиционный способ, им привычно пользуются при статистических анализах, хотя по удобству и простоте он может уступать другим методам.

Обычно ***F***-тест проводится путем сопоставления вычисленного значения ***F***-критерия с эталонным (табличным) показателем $F_{табл}$ для соответствующего уровня значимости. Если выполняется неравенство $F_{расч} \le F_{табл}$, то с уверенностью, например на 95%, можно утверждать, что рассматриваемая зависимость $y = b_0 + b_1x_1 + b_2x_2 + \ldots + b_kx_k$ является статистически значимой. Соответственно, и наоборот.

2. *Решение принимается на основе уровня значимости* α.

Для этого обратим внимание на представленные значения уровня значимости α (в интерпретации Excel это показатель ***p***). Если ***p***-значение больше, чем 0,05, то полученный результат нужно понимать как незначимый (для 95-процентной вероятности). В том случае, когда величина ***p*** оказывается меньше 0,05, то вывод такой – уравнение значимое с вероятностью 95%. Если же $p < 0{,}01$, то полученный результат является высоко значимым (т.е. степень риска ошибиться в нашем утверждении оказывается меньше 1%, или, что то же, степень надежности составляет 99%).

3. *Решение принимается на основе коэффициента детерминации* R^2.

В этом случае имеющуюся расчетную величину $R^2_{расч}$ (это то, что нам выдал Excel, см. рис.4.5) необходимо сравнить с табличными (критическими) значениями $R^2_{крит}$ для соответствующего уровня значимости (повторим еще

раз, обычно это 0,05). Если окажется, что $R^2_{расч} > R^2_{крит}$, то с упомянутой степенью вероятности (95%) можно утверждать, что анализируемая регрессия является значимой.

Теперь проанализируем наше уравнение с использованием рассмотренных статистических критериев.

1. Проведем проверку по F-критерию. Компьютерная распечатка выдала нам величину $F_{расч}$, равную 16,99 (см. лист Excel на рис.4.5). С учетом сделанных замечаний (стр.90) для анализа уравнения будем пользоваться величиной $F_{расч}$, обратной представленной Excel. Она составит 1:16,99=0,06. Отыщем по эталонной таблице (*Приложение 5*) критическую величину $F_{крит}$ при условии, что для числителя степень свободы $f_1 = n–k–1$= 20–2–1=17, а для знаменателя $f_2 = n–m$=20–1=19. Тогда для α=0,05 будем иметь $F_{крит} \approx$ 2,2. Понятно, что для всех рассмотренных вероятностей выполняется соотношение $F_{расч} < F_{крит}$, поэтому уверенно можно говорить о высокой степени адекватности анализируемого уравнения.

2. Теперь выполним проверку с использованием уровня значимости α (еще раз напомним, что Excel этот показатель именует как p). На рис.4.5 , где дано изображение листа Excel, находим позицию "Значимость F ". Там указана величина 8,84E-5, т.е. это число 8,84, перед которым стоит 5 нулей. Фактически можно признать, что α=0,000. Это говорит о том, что действительно обнаруживается устойчивая зависимость рассматриваемой функции y (величина продажи молока) от воздействующих факторов x_1 и x_2, т.е. объём реализации не является чисто случайной величиной. Правда, нам пока неизвестно, какие именно факторы (оба x_1 и x_2 или какой-то один из них) реально участвует в прогнозировании, но нам доподлинно понятно, что по крайней мере один из них влияет непременно.

3. Напомним, что по нашим расчетам коэффициент детерминации $R^2_{расч}$ составляет 0,67 или 67%. Таблица для тестирования на уровне значимости 5% в случае выборки n = 20 и числа переменных k = 2 дает критическое значение $R^2_{крит}$ = 0,297 (*Приложение 6*). Поскольку выполняется соотношение $R^2_{расч} > R^2_{крит}$, то с вероятностью 95% можно утверждать о наличии значимости данного уравнения регрессии.

Кстати заметим, что для наших обстоятельств (n = 20, k = 2) можно оценить критическое значение $R^2_{крит}$ для α=0,01 (высокая значимость). В этом случае $R^2_{крит}$ составляет 0,418, что, как видно, все равно остается мень-

ше расчетного показателя $R^2_{расч}$, т.е. 0,67. Из чего следует заключить, что обсуждаемое нами уравнение действительно характеризуется очень высокой степенью значимости.

Как видно, все три рассмотренных приема статистической проверки дают одинаковый результат. В этом примере мы воспользовались подобным разнообразием расчетов только с одной целью – дать представление о существующих методах такой проверки. На практике же нет нужды проводить статистическую оценку с использованием всех указанных вариантов. Вполне разумно (да и экономично) ограничиться каким-то одним методом. Каким именно? Более распространенным методом считается выполнение проверки по ***F***-критерию.

4.4.2. Проверка на адекватность коэффициентов регрессии

Итак, нами проведена проверка на значимость самого уравнения, т.е. мы понимаем, что существует взаимосвязь между параметром y и переменными x_i. Однако нам пока неясно, каково влияние на исследуемую функцию y конкретных факторов x_1 и x_2 (действуют ли они оба или только какой-то из них один). Поэтому предстоит определить значимость отдельных коэффициентов регрессии b_1 и b_2.

Проверку на адекватность коэффициентов регрессии рекомендуется проводить по следующим эквивалентным методам.

1. *Использование* ***t****-критерия*. Необходимые расчеты делает исполнительный Excel, который выдает соответствующую компьютерную распечатку с обозначением значений показателя t. Анализируемый коэффициент считается значимым, если его абсолютная величина превышает 2,00 (точнее 1,96), что соответствует уровню значимости 0,05. В нашем примере имеем для коэффициентов b_0, b_1 и b_2 следующие показатели критерия Стьюдента: t_{b0}=2,09; t_{b1}=2,59 и t_{b2}=2,57. Из представленного ряда следует, что значимыми оказываются все коэффициенты нашего уравнения.

2. *Использование уровня значимости*. В этом случае оценка проводится путем анализа показателя p, т.е. уровня значимости α. Коэффициент признается значимым, если рассчитанное для него p-значение (эти данные выдает Excel) меньше (или равно) 0,05 (т.е. для 95%-ной доверительной вероятности). Видим, что показатель p составляет для коэффициентов b_0, b_1 и b_2 следующие величины: p_{b0}= 0,05; p_{b1}= 0,02 и p_{b2}= 0,02.

Эти данные позволяют также заключить, что все рассмотренные коэффициенты статистически значимы. Иначе говоря, можно сделать вывод о неслучайном характере влияния всех изученных параметров.

Таким образом, проверка обоими методами дает вполне согласованные результаты. Поэтому в окончательном виде наше уравнение регрессии (для уровня значимости 0,05) следует записать так:

$$\hat{y} = 1{,}61+0{,}06\, x_1+0{,}07\, x_2.$$

4.5. Сравнительная оценка степени влияния факторов

При анализе полученного уравнения множественной регрессии закономерно встает вопрос, а какой фактор x_i из числа рассмотренных оказывает наибольшее влияние на исследуемый параметр y? К сожалению, исчерпывающего ответа на этот вопрос нет. Это связано с тем, что наличие возможной взаимосвязи между x-переменными (например, парное взаимодействие типа x_1x_2, тройное $x_1x_2x_3$ и т.д.) может сильно усложнить ситуацию. В результате станет принципиально невозможным выяснить, какая именно из переменных x_i в действительности отвечает за поведение параметра y.

Тем не менее, в статистике даются полезные рекомендации, позволяющие получить хотя бы оценочные представления по этому поводу. В качестве примера познакомимся с одним из таких методов – *сравнение стандартизованных коэффициентов регрессии*.

В общем случае все коэффициенты регрессии b_1, b_2 , …, b_k могут быть выражены в разных единицах измерения. Тем самым непосредственное их сравнение становится фактически некорректным, поскольку, скажем, формально м*е*ньший по величине коэффициент на деле может оказаться наиболее важным, чем б*о*льший. Короче говоря, в данной ситуации мы сталкиваемся с классической проблемой "сравнения кита и слона – кто кого поборет". *Стандартизованные коэффициенты регрессии* позволяют решить эту проблему за счет представления коэффициентов регрессии в некоторых кодированных единицах измерения.

Стандартизованный коэффициент регрессии вычисляется путем умножения коэффициента регрессии b_i на S_{Xi} и деления полученного произведения на S_y. Это означает, что каждый стандартный коэффициент регрессии измеряется как величина $b_i\, S_{Xi}/\, S_y$.

Применительно к нашему примеру получим следующие результаты (табл.4.4).

Т а б л и ц а 4.4

Стандартизованные коэффициенты регрессии

1. Стандартные отклонения		
Объем продажи S_y**= 1,06**	**Бабушки с внучками** S_{X1}**=8,26**	**Помощь Шарика** S_{X2}**=7,25**
2. Коэффициенты регрессии		
Бабушки с внучками b_1 **= 0,06**		**Помощь Шарика** b_2 **=0,07**
3. Стандартизованные коэффициенты регрессии		
Бабушки с внучками $b_1 S_{X1} / S_y$ **= 0,06×8,26/1,06=0,47**		**Помощь Шарика** $b_2 S_{X2} / S_y$ **=0,07×7,25/1,06=0,48**

Как видно, теперь мы можем вполне разумно сопоставлять полученные коэффициенты. Для обоих анализируемых факторов стандартизованные коэффициенты практически одинаковы.

Таким образом, приведенное сравнение абсолютных величин стандартизованных коэффициентов регрессии позволяет получить пусть и довольно грубое, но достаточно наглядное представление о важности рассматриваемых факторов. Еще раз напомним, что эти результаты не являются идеальными, поскольку не в полной мере отражают реальное влияние исследуемых переменных (мы оставляем без внимания факт возможного взаимодействия этих факторов, что может исказить первоначальную картину).

В целом же проведенный регрессионный анализ дает основание коту Матроскину по достоинству оценить коммерческий талант Шарика и задуматься о перспективах делового сотрудничества со своим приятелем из Простоквашино. Оказывает также влияние и конкретная категория покупателей (бабушки с внучками). Вместе с тем для Матроскина остаются поводы для творческих размышлений – он явно не принял во внимание какие-то иные факторы (вспомним про 33%, приходящихся на неучтенные причины), поскольку решил ограничиться рассмотрением более понятных и очевидных воздействий на результативность своего молочного бизнеса.

5. АНАЛИЗ «ХИ-КВАДРАТ»: ПОИСК ЗАКОНОМЕРНОСТЕЙ ДЛЯ КАЧЕСТВЕННЫХ ДАННЫХ

Когда не знаешь, что именно ты делаешь, делай это всегда тщательно.

(Правило Мэрфи)

Если качественные признаки не поддаются упорядочению, то использовать непараметрические способы уже нельзя. Единственный подсчет, который в этом случае можно выполнить, – это попытаться определить *частоты* проявления исследуемых признаков. Приходится прибегать к оценке наличия связи путем определения так называемого *хи-квадрата*.

Критерий *хи-квадрат* используют для проверки гипотез о качественных данных, представленных *не* числами, а категориями. Здесь принято оперировать подсчетом *частоты* (поскольку ранжирование или арифметические действия выполнять невозможно).

Критерий (тест) "хи-квадрат" основан на частотах, которые представляют собой количество единиц выборки, попадающую в ту или иную категорию. Суть показателя *хи-квадрат* (χ^2) – он измеряет *разницу* между *наблюдаемыми* (экспериментальными) *частотами* $f_Э$ и *ожидаемыми* (теоретическими) *частотами* f_T. Конкретно такой показатель рассчитывается как сумма квадратов разности этих частот, выраженная в долях частоты теоретической. Это утверждение можно записать следующим образом:

$$\chi^2 = \Sigma \frac{(f_Э - f_T)^2}{f_T}.$$

Использование такого статистического подхода возможно в разных обстоятельствах. Рассмотрим наиболее распространенные.

5.1. Комбинация: нынешние и прошлые события (критерий «хи-квадрат» соответствия)

Данный способ широко применяется в тех случаях, когда нужно прояснить ситуацию по поводу того, является ли наш *нынешний* опыт (выраженный в *частотах* или *процентах*) *типичным* по отношению к *прошлому* опыту (набор так называемых *опорных величин*). Такую ситуацию можно условно обозначить фразой "*Это* было недавно, а *то* было давно. Между ними есть соответствие?"

Тест "хи-квадрат" в отношении соответствия процентов используется для проверки гипотезы о том, что комбинация *наблюдаемых* частот или процентов (характеризующих одну качественную переменную) построена на данных из некоторой генеральной совокупности с уже известными значениями процентов (опорными величинами).

Можно сформулировать высказанные соображения и по-другому: те результаты, которые мы наблюдаем сейчас (фактические данные, т.е. наш нынешний опыт), на самом деле по характеру такие же, как и те, которые относятся к прошлым данным (опорным величинам). А это объясняется тем, что и те и другие относятся к одной и той же генеральной совокупности, просто они извлекались в разное время (сейчас и когда-то давно).

Ожидаемое значение частоты для каждой категории рассчитывается как произведение заданного *опорного* значения процента в генеральной совокупности на размер выборки ***n***. На основании имеющихся знаний о *наблюдаемой* частоте и частоте *ожидаемой* анализируемого события определяется собственно показатель *хи-квадрат*. *Расчетное* значение *хи-квадрат* затем сравнивают с *критическим* (табличным) показателем для соответствующего числа степеней свободы (определяется как *количество категорий минус единица*).

Если оказывается справедливым неравенство $\chi^2_{\text{расч}} > \chi^2_{\text{крит}}$, то с заданной вероятностью (или уровнем значимости) можно утверждать, что наблюдаемые частоты (наш опыт) значимо отличаются от тех, которые ожидаются исходя из известных нам опорных значений процентов (частот). Следовательно, обоснованно можно делать вывод о том, что *наблюдаемые выборочные проценты значимо отличаются от заданных опорных значений*.

Если имеем соотношение $\chi^2_{\text{расч}} \leq \chi^2_{\text{крит}}$, то наблюдаемые значения не очень-то отличаются от опорных показателей и, следовательно, *наши фактические результаты не имеют значимых отличий от заданных опорных значений*.

При выполнении такого анализа принято придерживаться следующего эмпирического правила: *ожидаемые частоты в каждой категории должны быть, по крайней мере, не меньше пяти* (поскольку критерий *хи-квадрат* остается приблизительной, а не совсем точной оценкой).

Анализ *критерия соответствия процентов (частот)* удобно выполнять, придерживаясь следующей схемы:

1. Имеются табличные данные частот для каждой категории одной качественной переменной. Обсуждаются следующие гипотезы:

а) частоты (проценты) нынешнего опыта равны набору известных, фиксированных опорных величин (из прошлого опыта);

б) частоты (проценты) нынешнего опыта не равны набору опорных величин (данных прошлого опыта).

2. *Ожидаемые частоты* вычисляются так: нужно для каждой категории умножить известное значение её доли в общем количестве (генеральной совокупности) на размер выборки ***n***.

При этом предполагается, что *а*) набор данных представляет собой случайную выборку из рассматриваемой генеральной совокупности и *б*) ожидается наличие, по крайней мере, пяти объектов в каждой из категорий.

3. *Анализ "хи-квадрат"* проводится с использованием уже упомянутого выражения

$$\chi^2 = \Sigma \frac{(f_Э - f_T)^2}{f_T}.$$

Степень свободы ***f*** рассчитывается так:

$$\boldsymbol{f = k} - 1,$$

где ***k*** – это число категорий, т.е. количество анализируемых параметров.

4. Интерпретация результата *теста "хи-квадрат"*: наличие значимой связи отмечается тогда, когда расчетное значение *"хи-квадрат"* больше табличного или критического (т.е. $\chi^2_{\text{расч}} > \chi^2_{\text{крит}}$), в противном случае значимой связи нет.

Теперь приступим к знакомству с анализом и, самое главное, выясним, как такой расчет можно выполнить с использованием компьютерной программы Excel.

Рассмотрим следующий пример.

Среди студентов, сдававших на первом курсе в летнюю сессию экзамен по математике, был проведен опрос с целью выяснить, какие факторы влияют на получение неудовлетворительной оценки. Число опрошенных студентов составляло 50 человек.

Наиболее часто упомянутыми причинами были указаны следующие:

1. Сам виноват, нужно было лучше заниматься.

2. Я знал, да, видите ли, профессор был не в духе.

3. К сожалению, не удалось списать.

4. Сказалось влияние роковых примет (достался билет № 13, повстречал черного кота, забыл надеть "счастливый" свитер и проч.).

Эти ответы можно условно разделить на следующие категории:

1. *Сам болван.*
2. *Вредный "препод".*
3. *Шпоры.*
4. *Черный кот.*

В табл.5.1 приведены данные о причинах получения "неудов" по математике за прошедшую сессию, а также указаны значения опорных величин, взятые из экзаменационных ведомостей по этому предмету за прошлые годы (по таким же категориям).

Как видно, по количественным показателям все анализируемые причины за прошедшую сессию формально отличаются от опорных значений. Однако это различие оказывается далеко неравноценным. Так, можно признать, что в категории самооценки ("Сам болван") фактические данные отличаются от соответствующих опорных величин относительно слабо (например, 57% по сравнению с 59% для прошлых сессий). В то же время по другим категориям относительное различие выглядит более заметным. Особенно бросается в глаза несоответствие по позиции "Шпоры".

Т а б л и ц а 5.1

Итоговые данные о причинах неудовлетворительной оценки по математике за прошедшую сессию и аналогичные данные (опорные) за прошлые годы

Причина	Наблюдаемые данные (за прошедшую сессию)		Опорные значения, % (ожидаемые данные)
	Частота	Процент от общего числа	

Сам болван	**28**	**57,0**	**59,0**
Вредный "препод"	**10**	**19,0**	**14,0**
Шпоры	**7**	**14,0**	**20,0**
Черный кот	**5**	**10,0**	**7,0**
Итого	**50**	**100**	**100**

Вопрос заключается в том, значима ли эта разница? Иначе говоря, могут ли полученные по итогам прошедшей сессии "неуды" рассматриваться как результат извлечения случайной выборки из генеральной совокупности, в которой проценты "неудов" соответствуют опорным величинам? Или еще по-другому: достаточно велика ли наблюдаемая разница, чтобы ее нельзя было объяснить только случайностью?

Тест *хи-квадрат* соответствия процентов позволит дать ответ на этот вопрос. Утвердительное заключение получим при условии, когда окажется справедливым соотношение $\chi^2_{\text{расч}} > \chi^2_{\text{крит}}$. Его нужно будет истолковать так: результаты нынешней сессии и результаты прошлых сессий отличаются между собой принципиально, поскольку различие между ними не носит случайного характера.

Если окажется справедливым неравенство $\chi^2_{\text{расч}} \le \chi^2_{\text{крит}}$, то с заданной вероятностью можно будет говорить о не значимости различия между анализируемыми результатами.

В табл.5.2 укажем частотные величины для обеих информационных позиций – текущие данные ("Наблюдение") и сведения за прошлые годы ("Ожидание"). Расчет частоты для графы "Ожидание" проведем путем умножения значений опорных величин процентов (59%, 14%, 20% и 7%) на размер выборки (n = 50). В результате получим следующие значения частот: 0,59×50=29,5; 0,14×50=7,0 и т.д. Заметим, что в итоговой строке для обеих колонок общая сумма частот одинакова – равна 50.

Т а б л и ц а 5.2

Наблюдаемые и ожидаемые данные (частоты) о причинах неудовлетворительных отметок

Причина	Наблюдение	Ожидание
Сам болван	**28**	**29,5**

Вредный "препод"	**10**	**7,0**
Шпоры	**7**	**10,0**
Черный кот	**5**	**3,5**
Итого	**50**	**50,0**

Эти данные и будем использовать для решения вопроса о значимом соответствии (или несоответствии) фактических и ожидаемых результатов. Воспользуемся для этого теми возможностями, которые предоставляет приложение Excel. Напомним, что нам для анализа нужно располагать величинами $\chi^2_{\text{расч}}$ и $\chi^2_{\text{крит}}$. Все эти характеристики вычисляются с помощью приложения Excel.

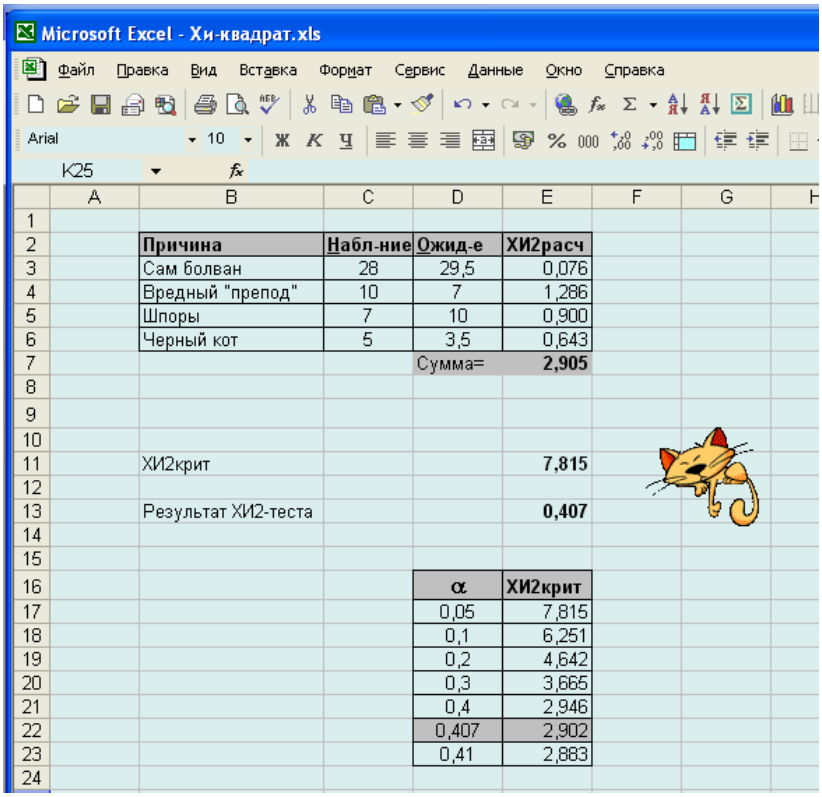
Пояснение. Вообще-то значения $\chi^2_{\text{крит}}$, как обычно это делается при статистическом анализе, извлекаются из специальных таблиц, содержащих заранее рассчитанные эталонные значения этой характеристики. Однако в нашем случае используем возможности Excel, поскольку подобную услугу он способен оказать совершенно элементарно.

1. Откроем лист Excel и составим нашу таблицу с имеющими данными (рис. 5.1). Пусть они будут находиться в диапазоне ячеек (вместе с названиями) B2:D6. Пристроим к таблице еще одну графу (E2:E6), в которой, помимо заголовка, будут находиться расчетные значения *хи-квадрат*, вычисленные для каждой строки (т.е. для каждого анализируемого фактора).

2. Расчет проведем по известной уже формуле, запись которой представлена в виде:

$$\mathbf{ХИ2_{расч}=\Sigma(f_Э-f_Т)\text{^}2/\ f_Т,}$$

где $f_Э$ и $f_Т$ − соответственно экспериментальные (наблюдаемые) и теоретические (ожидаемые) значения частот.



	A	B	C	D	E	F	G
1							
2		**Причина**	**Набл-ние**	**Ожид-е**	**ХИ2расч**		
3		Сам болван	28	29,5	0,076		
4		Вредный "препод"	10	7	1,286		
5		Шпоры	7	10	0,900		
6		Черный кот	5	3,5	0,643		
7				Сумма=	**2,905**		
8							
9							
10							
11		ХИ2крит			**7,815**		
12							
13		Результат ХИ2-теста			**0,407**		
14							
15							
16				**α**	**ХИ2крит**		
17				0,05	7,815		
18				0,1	6,251		
19				0,2	4,642		
20				0,3	3,665		
21				0,4	2,946		
22				0,407	2,902		
23				0,41	2,883		
24							

Рис.5.1. Фрагмент рабочего листа Excel с исходными данными и результатами анализа *хи-квадрат*

Чтобы выполнить расчет для данных первой строки, выделим ячейку E3 и в строке формул запишем =(C3-D3)^2/ D3. Полученный результат расчета появится в этой ячейке. С округлением до третьего знака это составит 0,076. Аналогичные вычисления проделаем для остальных позиций. Для этого вновь выделим ячейку E3 и протянем ***Маркер заполнения*** (маленький квадратик в правом нижнем углу) вдоль всей графы вниз – во всех соответствующих ячейках будут содержаться готовые расчетные значения *хи-квадрат*.

Просуммируем эти данные, получим величину 2,905. Это и есть наш искомый $\chi^2_{\text{расч}}$.

3. Теперь займемся вычислением показателя $\chi^2_{\text{крит}}$. Для этого применим функцию **ХИ2ОБР**. Для ее запуска предназначена специальная программа. Воспользуемся ***Мастером функций***.

Поступим следующим образом:

– выделим ту ячейку, в которой должен находиться получаемый результат;

– активизируем ***Мастер функций*** кнопкой f_x;

– в появившемся диалоговом окне выберем нужную категорию из имеющегося списка и укажем опцию ***Статистические***;

– затем отыщем собственно нужную нам функцию ***Хи2обр***, после чего нажмем на кнопку ***ОК***.

4. На экране появится диалоговое окно для ввода параметров, необходимых для вычисления критического (табличного) значения *хи-квадрата* (рис.5.2). В первом текстовом поле ввода (***Вероятность***) укажем выбранную величину уровня значимости α. Примем традиционный показатель степени риска, равный 0,05.

5. Во втором поле ввода (***Степени свободы***) запишем число степеней свободы. Поскольку в нашем примере фигурируют четыре компонента (причины "неудов"), то число степеней свободы составит: $f = k - 1 = 4 - 1 = 3$.

После нажатия на кнопку ***ОК*** в выбранной нами ранее ячейке (E11) появится значение $\chi^2_{крит}$, равное 7,815 (после надлежащих округлений).

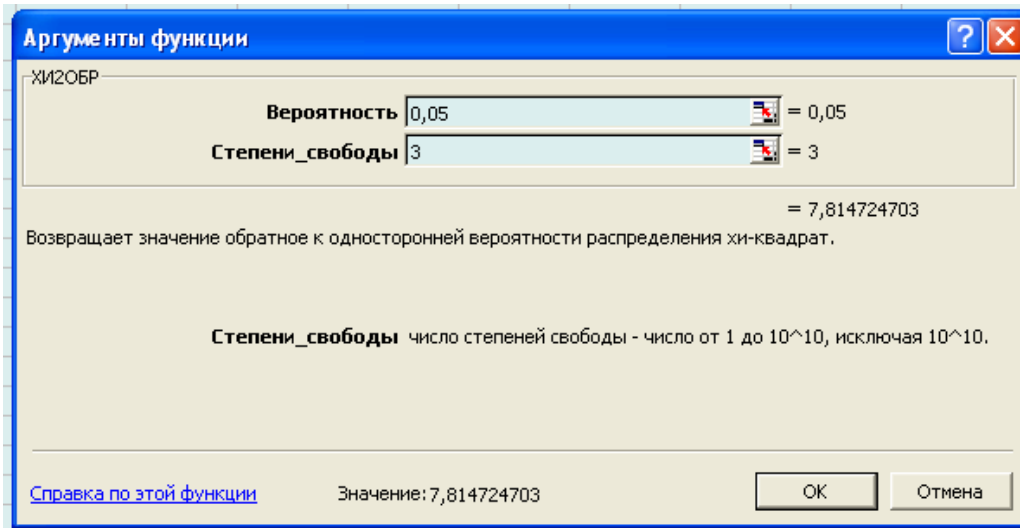


Рис.5.2. Диалоговое окно ввода параметров для определения критического (табличного) значения *хи-квадрат*

Вот с этим-то числом нам и нужно будет затем сравнивать расчетное значение $\chi^2_{расч}$. Поскольку выполняется соотношение $\chi^2_{расч} \le \chi^2_{крит}$ (ибо 2,905 < 7,815), то с вероятностью 95% можно утверждать, что наблюдаемые (фактические) показатели незначимо отличаются от ожидаемых (опорных) значений.

Анализ *хи-квадрат* в режиме Excel можно выполнить и по-другому, с использованием так называемого *хи-теста*. Функция **ХИ2ТЕСТ** позволяет определить вероятность того, является ли различие между наблюдаемыми и ожидаемыми значениями статистически значимым результатом.

Покажем это на нашем примере.

6. Для этого вновь действуем с помощью ***Мастера функций***:

– выделяем ячейку (допустим E13), в которой должен находиться получаемый результат;

– активизируем ***Мастер функций***;

– в диалоговом окне выбираем нужную категорию и указываем опцию ***Статистические***;

– отыскиваем функцию ***Хи2тест***, после чего нажимаем на кнопку ***ОК***.

В появившемся диалоговом окне (рис.5.3) нужно заполнить текстовые поля, в которых следует указать имеющиеся данные, относящиеся к фактическим и ожидаемым результатам. Напомним, эти данные занимают соответственно ячейки C3:C6 и D3:D6.

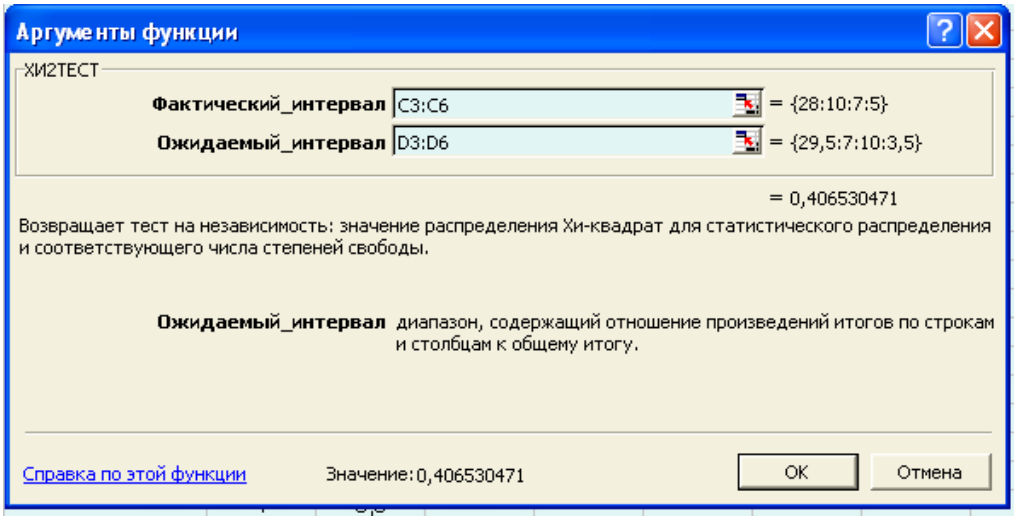


Рис.5.3. Диалоговое окно ввода параметров для определения расчетного значения *хи-квадрат*

Кстати, после введения интервальных ячеек справа от каждого поля ввода в скобках будут перечислены те табличные значения, которые содержались в соответствующих столбцах (рис.5.3). Там же в окне можно будет

прочитать и полученное расчетное значение *уровня значимости*, равное 0,406530471. А после нажатия на клавишу ***ОК*** этот результат будет помещен в выделенную нами ячейку.

Проведем округление полученного результата до третьего знака после запятой и в окончательном виде получим 0,407. Теперь попытаемся оценить полученные данные.

Указанное число показывает, что гипотеза о том, что результаты нынешней сессии отличаются от итогов прошлых лет, высказывается с риском допустить ошибку на 40,7%. И напротив, почти с вероятностью 60% можно говорить о том, что различие между этими данными несущественное.

Как же следует толковать данные анализа *хи-квадрат*, исполненные обоими способами (сравнением $\chi^2_{\text{расч}}$ и $\chi^2_{\text{крит}}$, а также применением функции *хи2-тест*)? Покажем, что оба подхода идентичны.

1. Нами сделано заключение о статистической неразличимости наблюдаемых и ожидаемых результатов на основании сопоставления значений $\chi^2_{\text{расч}}$ (2,905) и $\chi^2_{\text{крит}}$(7,815). Напомним, что этот вывод был сделан для уровня значимости α=0,05 (т.е. для 5-процентной степени риска). Теперь попытаемся выяснить, при каких же условиях можно отважиться на утверждение, что экзаменационные данные нынешние и прошлые (по характеру влияния на их итоги рассматриваемых факторов) все-таки разнятся. Иными словами, полагать, что с точки зрения статистического подхода эти данные являются извлечением не из одной и той же генеральной совокупности, а принадлежат совершенно различным массивам.

Для этого, используя функцию **ХИ2ОБР,** рассчитаем значения $\chi^2_{\text{крит}}$ для различных уровней значимости, постепенно повышая вероятность допустить ошибочный прогноз (увеличивая α). На рабочем листе Excel (рис.5.1) в виде списка приведены полученные значения $\chi^2_{\text{крит}}$ для α, равного соответственно 0,05; 0,1; 0,2 и т.д. Закончим же расчет для случая α=0,407 и 0,41. Почему именно эти числа, станет сейчас понятным.

Наше расчетное значение $\chi^2_{\text{расч}}$ (2,905) окажется превышающим $\chi^2_{\text{крит}}$ (2,902), когда α будет *больше* 0,407. Например, для α=0,41 уже можно определенно говорить, что условие $\chi^2_{\text{расч}} > \chi^2_{\text{крит}}$ (2,905* > 2,883) выполняется и допустимо утверждение, что обе рассматриваемые совокупности являются различными.

---

* Числа 2,902 и 2,905 - это фактически одно и то же, различие обусловлено некоторым искажением при выполнении операции округления.

2. Теперь дадим оценку только что сделанному заявлению. Прелесть статистики состоит в том, что она любое утверждение всегда дает с определенной гарантией надежности, т.е. страхуется от проявления возможных случайностей (погрешностей). Совершенно недостаточно высказать какое-то соображение. Важно обязательно также определить, с какой степенью вероятности (или уровнем риска впасть в ошибку) оно формулируется.

Когда мы заявили, что влияние рассматриваемых факторов на итоги прошедшей сессии и сессий прошлых лет различаются, то сделали это с риском оказаться неправыми почти на 41%! Совершенно чудовищная степень ошибочности утверждения! Кто всерьез примет в расчет такое мало обоснованное соображение?

Поэтому в ситуациях, когда мы должны высказывать суждения с достаточной степенью надежности (обычно при α=0,05, а еще лучше 0,01), величина порогового (критического) значения χ^2 имеет очевидную тенденцию к возрастанию. А это означает, что при разумном объеме единиц наблюдения (в данном случае это студенты, большие знатоки математической науки) мы лишь можем говорить о незначимости рассматриваемых итогов. Чтобы все-таки обнаружить подобное возможное различие, следовало было бы провести более масштабное по охвату обследование. Однако можно тихо утешиться тем обстоятельством, что проделать всю эту процедуру весьма проблематично вследствие недостаточного числа (смеем надеяться!) физически наличествующих двоечников.

Итак, резюме. Для обсуждаемого примера можно заключить, что "неуды" по математике, полученные в прошедшую сессию, по характеру причин (в интерпретации самих студентов) соответствуют тем же показателям, что случались и в прошлые годы. Имеющиеся расхождения обусловлены только лишь случайностью (для выборки размером 50). У нас нет убедительных причин полагать, что воздействующие прискорбные факторы как-то принципиально изменились (т.е. как было раньше, так и осталось нынче) и повлияли на результативность сдачи экзамена. По-прежнему доминирующей причиной остается собственная нерадивость студентов, а изменения остальных факторов вполне укладываются в границы случайных колебаний. Так что в этом отношении у деканата нет повода для беспокойных раздумий.

5.2. О коэффициентах взаимной сопряженности

На основе *хи-квадрата* принято также оценивать показатели *степени тесноты связи – коэффициенты взаимной сопряженности* К.Пирсона и А.Чупрова.

Коэффициент Пирсона рассчитывается по формуле:

$$K_{П} = \sqrt{\frac{\chi^2}{n + \chi^2}},$$

где χ^2 – расчетное значение *хи-квадрата*, n – общее число наблюдений (объем выборки).

Коэффициент Чупрова позволяет учесть число групп по каждому признаку и определяется следующим образом:

$$K_{Ч} = \sqrt{\frac{\chi^2}{n\sqrt{(k_1 - 1)(k_2 - 1)}}},$$

где k_1 и k_2 – соответственно число значений (групп) для первого и второго признаков или, по-другому, число строк и столбцов в таблице, а n – общее число наблюдений.

Попробуем выполнить такие расчеты для нашего примера.

$$K_{П} = \sqrt{\frac{\chi^2}{n + \chi^2}} = \sqrt{\frac{2{,}905}{50 + 2{,}905}} = 0{,}234;$$

$$K_{Ч} = \sqrt{\frac{\chi^2}{n\sqrt{(k_1 - 1)(k_2 - 1)}}} = \sqrt{\frac{2{,}905}{50 \cdot \sqrt{(3-1)(2-1)}}} = 0{,}205.$$

Как видно, расчет обоих коэффициентов дает весьма малые величины, что свидетельствует об отсутствии связи между исследуемыми характеристиками. Это же подтверждают и оценки по таблице Чеддока – рассчитанные коэффициенты (по модулю меньше 0,3) говорят об отсутствии корреляционной связи. Иначе говоря, использование и этих коэффициентов подтверждает ранее высказанное соображение – анализируемая ситуация по своим параметрам соответствует опорным (ожидаемым) показателям и посему не требует введения каких-либо корректировок.

5.3. Проверка взаимосвязи между двумя качественными переменными (критерий «хи-квадрат» независимости)

Возможны ситуации, когда имеются две качественные переменные, характеризующие события, не связанные с временным фактором. После изучения каждой из них *отдельно* с помощью анализа частот (или процентов) может возникнуть вопрос о наличии *связи* между ними.

Считается, что две качественные переменные являются *независимыми*, если знание значения одной переменной не помогает предсказать значение другой.

Представим себе, что ваша фирма разработала технологию гальванического покрытия никелем стальных деталей автомобильного кузова. В среднем процент брака, связанного с отслаиванием покрытия, составляет 3,1%. Однако когда работает технолог г-н Безенчук, размер брака достигает 11,2%. В этом случае знание значения одной переменной (имя конкретного технолога) помогает спрогнозировать значение другой переменной (объем брака определенного типа), поскольку 3,1% и 11,2% различаются между собой. Появление брака более вероятно во время работы г-на Безенчука и менее вероятно, когда работает кто-то другой. Следовательно, эти две переменные *не являются независимыми*.

Использование критерия "*хи-квадрат*" позволяет решить вопрос о том, являются ли рассматриваемые качественные совокупности зависимыми или же независимыми друг от друга. В этом случае применяется так называемый *критерий "хи-квадрат" независимости,* который устанавливает наличие (или отсутствие) связи между двумя качественными переменными. Для такого анализа используется таблица частот, которые можно было бы ожидать в том случае, если переменные оказались бы независимыми.

В общем случае *критерий "хи-квадрат" независимости* принято представлять в виде такой схемы:

1. Имеются исходные данные в форме табличного списка частот всех комбинаций категорий двух качественных переменных. Обсуждаются следующие гипотезы:

а) две переменные не зависят одна от другой;

б) две переменные связаны, они не являются независимыми друг от друга.

2. Составляется *таблица ожидаемых частот*. Для их расчета частоту одной категории (результат эксперимента) следует умножить на частоту дру-

гой категории (также экспериментальный показатель) и полученное произведение поделить на общий объем выборки ***n***:

$$\text{Ожидаемая частота } f_{Ож\,(Т)} = \frac{\text{Частота категории } f_{Э1} \times \text{Частота категории } f_{Э2}}{\text{Общий объем выборки } n}$$

или более компактно в символьной форме: $f_{ОЖ(Т)} = \dfrac{f_{Э1} \cdot f_{Э2}}{n}$.

При этом считается, что *а*) набор данных представляет собой случайную выборку из рассматриваемой генеральной совокупности и *б*) для каждой комбинации категорий ожидаемая частота, по крайней мере, не меньше пяти.

3. Далее проводится *анализ "хи-квадрат"*, расчет выполняется с использованием знакомого выражения:

$$\chi^2 = \Sigma \frac{(f_Э - f_Т)^2}{f_Т}.$$

Степень свободы вычисляется следующим образом:

$$f = (k_1 - 1) \times (k_2 - 1),$$

где k_1 и k_2 – число категорий соответственно для первой и второй переменной.

4. Результат *теста "хи-квадрат"* трактуется так: наличие значимой связи проявляется тогда, когда расчетное значение *"хи-квадрат"* больше критического (т.е. $\chi^2_{расч} > \chi^2_{крит}$), в противном случае значимой связи нет.

Более подробно познакомимся с этим видом статистического анализа, для чего рассмотрим следующий пример.

Кот Матроскин, занявшись молочным бизнесом, решил провести маркетинговое исследование, чтобы уяснить, какой вид молочной продукции предпочитают те или иные покупатели. Для каждой покупки фиксировались две качественные переменные – вид продукции и тип покупателя. В качестве продаваемой молочной продукции фигурировали молоко, сметана и творог. Покупателей Матроскин условно разделил на две категории – практичные и импульсивные. К первым он отнес тех покупателей, которые идут на рынок

уже с четко сформулированным намерением относительно того, что купить и сколько именно. Вторую же категорию составили покупатели, которые решение принимают на месте, непосредственно перед покупкой.

Полученные данные статистического опроса аккуратный кот Матроскин представил в табличной форме (табл.5.3), в которой для каждого вида молочной продукции указал количество совершаемых покупок тем или иным покупателем (т.е. привел фактическую частоту).

Необходимо дать заключение по итогам статистической проверки по критерию "хи-квадрат", т.е. сформулировать вывод и пояснить результат с практической точки зрения (какую рыночную стратегию должен избрать кот Матроскин и, следовательно, на какого покупателя и на какой вид молочной продукции ему надлежит ориентироваться).

Решение этой задачи вновь проделаем в двух вариантах – консервативным способом ("вручную") и компьютерным.

Таблица 5.3

Результаты опроса о перспективах молочного бизнеса

Вид молочной продукции	Частота предпочтений	
	Практичный покупатель	Импульсивный покупатель
Молоко	38	15
Сметана	24	31
Творог	18	27

Но сначала таблицу с исходными данными дополним. Для этого введем итоговые строку и столбец и их заполним, выполнив несложные расчеты (табл.5.4).

Чисто визуально трудно ответить, есть ли взаимосвязь между этими признаками (разными категориями покупателей и видами молочной продукции).

Таблица 5.4

Дополненные данные по результатам опроса о перспективах молочного бизнеса

Вид молочной продукции	Частота предпочтений		Итого
	Практичный покупатель	Импульсивный покупатель	
Молоко	38	15	53
Сметана	24	31	55
Творог	18	27	45

Итого	**80**	**73**	**153**

Поэтому необходимо дать анализ распределения частот в таблице по строкам и графам.

Будем исходить из следующего постулата. Если признак, положенный в основу группировки по строкам (*вид молочной продукции*), *не* зависит от признака, положенного в основу группировки по столбцам (*тип покупателя*), то в *каждой* строке (столбце) распределение частот должно быть пропорционально распределению их в *итоговой* строке (столбце). Такое распределение можно рассматривать как *теоретическое* (ожидаемое), частоты которого рассчитаны в предположении *отсутствия* связи между изучаемыми совокупностями.

Рассчитаем *ожидаемые* частоты *внутри* таблицы пропорционально распределению частот в *итоговой* строке.

Так, *молоко* как один из видов молочной продукции в зависимости от поведения посетителей рынка по частоте попадания в категории "Практичный покупатель" и "Импульсивный покупатель" имеет следующие показатели:

$$f_{11} = \frac{53 \cdot 80}{153} = 27{,}7 \quad f_{12} = \frac{53 \cdot 73}{153} = 25{,}3 .$$

Для второй стоки, т.е. для категории *сметана*, эти показатели имеют уже такие значения:

$$f_{21} = \frac{55 \cdot 80}{153} = 28{,}8 ; \quad f_{22} = \frac{55 \cdot 73}{153} = 26{,}2 .$$

Для третьей строки (категория *творог*):

$$f_{31} = \frac{45 \cdot 80}{153} = 23{,}5 ; \quad f_{32} = \frac{45 \cdot 73}{153} = 21{,}5 .$$

Полученные результаты (вычисленные значения частот) поместим в табл.5.5 .

Т а б л и ц а 5.5

Данные о перспективах молочного бизнеса с учетом ожидаемых частот

Вид молочной	**Ожидаемая частота предпочтений**	**Итого**

продукции	Практичный покупатель	Импульсивный покупатель	
Молоко Сметана Творог	27,7 28,8 23,5	25,3 26,2 21,5	53 55 45
Итого	80	73	153

Расчетное значение критерия *хи-квадрат* определим по формуле:

$$\chi^2 = \sum_{i=1}^{k_{21}} \sum_{j=1}^{k_1} \frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*},$$

где f_{ij} и f_{ij}^* – соответственно фактические и теоретические (ожидаемые) частоты в ***i***-й строке и ***j***-го столбца; k_1 и k_2 – соответственно число категорий в строках и столбцах таблицы.

Выполним соответствующие расчеты:

$$\chi^2_{расч} = \frac{(38-27{,}7)^2}{27{,}7} + \frac{(15-25{,}3)^2}{25{,}3} + \frac{(24-28{,}8)^2}{28{,}8} + \frac{(31-26{,}2)^2}{26{,}2} +$$
$$+ \frac{(18-23{,}5)^2}{23{,}5} + \frac{(27-21{,}5)^2}{21{,}5} = 12{,}4$$

Далее полагается сравнить расчетное значение $\chi^2_{\text{расч}}$ с табличным показателем (обычно для уровня значимости 0,05 или 0,01). В рассматриваемом примере число степеней свободы равно трем, т.е. f= (3–1) (2–1)=2[*]. При α=0,05 табличное значение $\chi^2_{\text{табл}}$ при f= 2 составляет 5,99, а для α=0,01 соответственно 9,21 (*Приложение 7*). Поскольку $\chi^2_{\text{расч}} > \chi^2_{\text{табл}}$, то с уверенностью на 95% и даже 99% можно утверждать, что влияние психологического типа покупателя очевидным образом сказывается на результатах коммерческой деятельности кота Матроскина. Ему, как видно, есть над чем поразмышлять.

Теперь посмотрим, что нам покажет трудяга Excel.

Прежде всего, перенесем табл.5.4 и 5.5 в рабочий лист Excel (рис.5.4). При этом в ячейке А22 запишем "ХИ2крит", а соседние ячейки В22 и С22 зарезервируем за численными значениями $\chi^2_{\text{крит}}$. Считать будем для двух зна-

[*] В данном случае и частота и степень свободы обозначены одним и тем же буквенным символом ***f***.

чений уровня значимости – 0,05 и 0,01 (их заголовки разместим в ячейках B21 и C21). Кроме того, в ячейках A30 и A32 запишем "ХИ2расч" и "Рез-т ХИ2-тест", а соседние ячейки B39 и B32 подготовим для будущих итоговых расчетов.

После этого приступим собственно к самой работе в компьютерном варианте.

	A	B	C	D	E
3				Табл.5.4	
4	Виды молочной	Практичный покупатель	Импульсивный покупатель	Итого	
5					
6	Молоко	38	15	53	
7	Сметана	24	31	55	
8	Творог	18	27	45	
9	Итого	80	73	153	
10					
11		27,7124183	25,2875817		
12		28,75816993	26,24183007		
13		23,52941176	21,47058824		
14				Табл.5.5	
15	Виды молочной	Практичный покупатель	Импульсивный покупатель	Итого	
16	Молоко	27,7	25,3	53	
17	Сметана	28,8	26,2	55	
18	Творог	23,5	21,5	45	
19	Итого	80	73	153	
20					
21		Ур. знач. 0,05	Ур. знач. 0,01		
22	ХИ2крит	5,991	9,210		
23					
24		3,8			
25		0,8			
26		1,3			
27		4,2			
28		0,9			
29		1,4			
30	ХИ2расч =	12,4			
31					
32	Рез-т ХИ2-тест	0,002			
33					

Рис.5.4. Лист Excel с результатами расчета критерия *хи-квадрат*

1. Для определения показателя $\chi^2_{\text{крит}}$ применим функцию **ХИ2ОБР.** Воспользуемся ***Мастером функций***, а затем командами ***Статистические/ Хи2обр.***

При заполнении диалогового окна укажем следующие параметры: для α = 0,05 и 0,01, а для степени свободы – 3.

После исполнения всех манипуляций и необходимых округлений в ячейках B22 и C22 будут содержаться следующие результаты: 7,815 и 11,345.

Затем произведем необходимые подсчеты ожидаемых частот. Используем уже знакомое выражение:

$$f_{ОЖ(Т)} = \frac{f_{Э1} \cdot f_{Э2}}{n}$$

2. Здесь поступим следующим образом. Вычисленные значения будем помещать в диапазоне ячеек B11:C13. Запишем формулу вычисления ожидаемых частот, которую затем станем копировать для заполнения всей таблицы. Будем использовать знак $ для задания "абсолютного адреса". Так, для расчета первого ожидаемого значения частоты используем выражение =B$9*$D6/$D$9 и получим 27,7124183 (с округлением 27,7).

3. Чтобы теперь получить остальные значения ожидаемых частот, проделаем следующее. Выделим ячейку B11 и появившийся маркер заполнения протянем вниз, захватывая ячейки B12 и B13. Тот же час в ячейках окажутся рассчитанные значения частот. Если теперь эти ячейки последовательно выделять и протягивать вправо, то в диапазоне C11:C13 появятся остальные показатели.

А теперь посмотрим на эти результаты и на скопированную нами табл.5.5 с ожидаемыми частотами. Что-то ужасно знакомое! С учетом необходимых округлений это же полная копия.

Теперь мы наглядно представляем, насколько легко Excel справляется с расчетами, над которыми нам перед этим (вспомним ручной счет) пришлось основательно потрудиться.

Продолжим расчеты.

4. Анализ *хи-квадрат* выполним с помощью функции **ХИ2ТЕСТ**. Действием уже привычным образом, используя следующие команды: ***Мастер функций/ Статистические / Хи2тест***.
Ячейку B32 выделим для **ХИ2ТЕСТ**.

5. При заполнении диалогового окна (рис.5.5) в текстовом поле ***Фактический интервал*** укажем адрес ячеек B6:C8, в которых находятся экспериментальные данные по частотам (табл.5.4). Соответственно в текстовом поле ***Ожидаемый интервал*** укажем диапазон B16:C18, содержимое которого отражает теоретические значения частот (табл.5.5).

В окончательном виде в ячейке B32 будет находиться следующий показатель – 0,002.

Как же следует трактовать полученный результат? Тезис о независимости обсуждаемых параметров (вид молочной продукции и психологический

тип покупателя) можно было бы принять, если бы уровень значимости α был бы меньше 0,002. Но для 95-процентной вероятности и даже 99-процентной установленные значения α (0,05 и 0,01) превышают 0,002. Это говорит о высокой степени значимости и, следовательно, указанные две качественные переменные являются вполне зависимыми друг от друга.

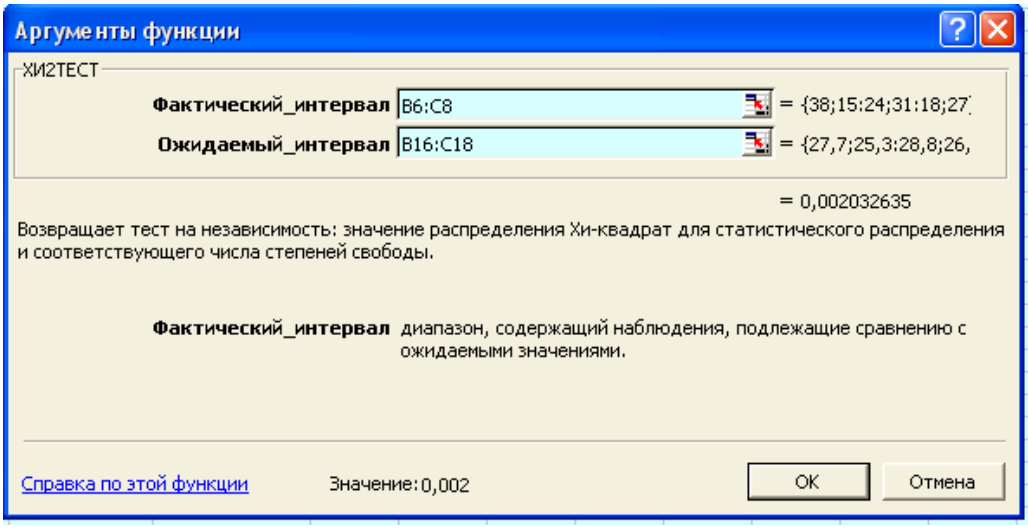


Рис.5.5. Диалоговое окно ввода параметров

И еще. Вспомним, что вывод о значимости связи между сопоставляемыми переменными можно сделать также на основе сравнения значений $\chi^2_{\text{расч}}$ и $\chi^2_{\text{табл}}$. Табличные значения у нас уже есть, это 7,815 и 11,345 (для уровней значимости 0,05 и 0,01). Теперь рассчитаем $\chi^2_{\text{расч}}$, для этого по формуле $\chi^2 = \dfrac{(f_Э - f_T)^2}{f_T}$ для каждой комбинации *наблюдаемых* (экспериментальных) $f_Э$ и *ожидаемых* (теоретических) *частот* f_T вычислим текущие значения χ^2, а затем их просуммируем. Результат приведен в виде списка на рис.5.4 и он, как и в случае ручного счета, равен 12,4 (ячейка B30). Ну а дальше знакомые процедуры – сравнение значений $\chi^2_{\text{расч}}$ (12,4) и $\chi^2_{\text{табл}}$ (7,815 и 11,345) указывает на то, что анализируемые качественные переменные не являются независимыми (мы это утверждаем с риском ошибиться на 5 и даже 1%).

Как видно, и ручной и компьютерный расчеты приводят нас к одному и тому же статистическому выводу – значимая связь между двумя рассматриваемыми качественными совокупностями имеет место быть.

Таким образом, как мы и утверждали по итогам ручного счета, коту Матроскину надлежит внимательно продумать свою дальнейшую коммерче-

скую стратегию – продаваемая продукция существенного зависит от того, кто ее покупает. Вот так! Причем наиболее заметно это проявляется по поводу торговли молоком. Очевидно, что свежее молоко предпочитают главным образом покупатели основательные, хорошо обдумывающие свой поход на рынок. В тоже время импульсивные визитеры эту продукцию заметно игнорируют, более полагаясь на сметану и творог.

Вот какого рода соображения можно высказать на основании выполненного анализа.

6. СТАТИСТИЧЕСКИЕ МЕТОДЫ ИЗУЧЕНИЯ ДИНАМИЧЕСКИХ ПРОЦЕССОВ

Прогнозы бывают трех видов: верные, неверные и научные.

(Гаврила Увеков)

Чтобы прослыть ясновидцем, предсказывай будущее на сто лет вперед. Чтобы прослыть глупцом, предсказывай его на завтра.

(Неизв. автор)

Временными (хронологическими) рядами или *рядами динамики* называются такие ряды, в которых статистические данные находятся в функциональной зависимости от времени.

В случае динамических рядов сама *последовательность* наблюдений несет в себе важную информацию. Так, чтобы охарактеризовать какую-то совокупность данных *в целом*, нам уже недостаточно знать лишь типичное значение этих данных (например, среднее арифметическое или стандартное отклонение). В данной ситуации желательно знать, что, скорее всего*, произойдет дальше*. Тем самым подобный прогноз должен *экстраполировать* ближайшее поведение исследуемой системы с точки зрения её функционирования в *прошлом.*

6.1. Понятие о статистических рядах динамики

Таким образом, главная цель анализа временн*ы*х рядов заключается в *создании прогнозов*, т.е. *предсказание будущего*. Эти прогнозы основываются на той или иной модели (ее называют *математической моделью* или *процессом*). Модель представляет собой систему уравнений, которая позволяет получить некий набор *искусственных* совокупностей данных, относящихся к категории временных рядов. Прогноз позволяет получить ожидаемое (т.е. среднее) значение будущего поведения оцениваемой модели.

Подобно всем оценкам, прогноз обычно не в полной мере соответствует действительности. *Границами прогноза* являются доверительные границы прогноза (если используемая модель позволяет их определить). Если модель

корректна по отношению к исследуемым данным, то будущее наблюдение с вероятностью, например, 95% попадает в эти границы.

Следовательно, *динамическими рядами* называются статистические данные, отражающие развитие исследуемого процесса во времени.

В каждом ряду динамики содержатся два основных элемента:

– *показатель времени* τ;

– соответствующие им *уровни развития* изучаемого процесса y.

В качестве *показателя времени* в рядах динамики выступают

– либо определенные *даты* (моменты) времени;

– либо отдельные *периоды* (годы, кварталы, месяцы, сутки).

Статистические показатели, характеризующие изучаемый объект, называют *уровнями ряда*. Уровни отображают количественную оценку (меру) развития исследуемого процесса во времени.

По форме выражения уровни могут быть *абсолютными, относительными* или *средними* величинами. При этом они могут отражать состояние процесса

− на определенный *момент времени* (на начало месяца, квартала, года);

− за определенные *интервалы времени* (за сутки, месяц, год и т.п.).

Соответственно по фактору времени принято различать *моментные* и *интервальные* динамические (временные) ряды (рис.6.1).

Рис.6.1. Схематическое изображение рядов динамики по фактору времени

Отличительная особенность интервальных рядов динамики абсолютных величин − можно *суммировать* их уровни, поскольку они *не* содержат *повторного счета*. Тем самым можно суммировать уровни как за более ко-

роткий промежуток времени (сутки, недели, месяц), так и за более длительный (квартал, год). В результате суммирования уровней интервального динамического ряда получаются так называемые *накопленные итоги*, которые имеют вполне реальное содержание.

Пример интервального ряда динамики – скажем, выпуск университетом специалистов по годам.

Вместе с тем моментным рядом динамики может служить, допустим, число студентов в университете. Уровни данного ряда − это обобщенные итоги учета числа студентов по состоянию на определенную дату. Ею может быть конец или начало соответствующего учебного года. При этом отдельные уровни моментного ряда динамики содержат элементы так называемого *повторного счета*. Суть сказанного состоит в следующем: большая часть студентов, учитываемая, например, в прошедшем учебном году, естественно, наличествует (за вычетом выпускников и отчисленных) и в настоящее время, являясь единицами совокупности и в текущем учебном году. Вот почему суммирование уровней моментных рядов динамики (в отличие от интервальных) становится процедурой, лишенной очевидного смысла.

6.2. Изучение основной тенденции развития

Мы рассмотрели наиболее используемые статистические характеристики, которые применяются для количественной оценки динамических рядов. Теперь основное внимание уделим тому, как на основании анализа временн*о*го ряда можно прогнозировать развитие событий в будущем.

Важным направлением в исследовании закономерностей экономических процессов является изучение *общей тенденции развития (тренда)*.

Изменения уровней временных рядов обусловливаются влиянием на изучаемый процесс различных факторов. В общем случае они неоднородны по силе, направлению и времени воздействия. Принято выделять так называемые *систематическую* ***Ст*** и *случайную* ***Сл*** составляющие. При этом в зависимости от формы разложения динамического ряда различают *аддитивную* и *мультипликативную* модели. В первом случае исходные данные динамического ряда ***ИД*** описывается в виде суммы этих показателей (т.е. выражением ***ИД*** = ***Ст*** + ***Сл***) , во втором – их произведением (***ИД*** = ***Ст*** · ***Сл***).

В свою очередь систематическая составляющая ***Ст*** является интегральной характеристикой, поскольку отражает влияние нескольких факторов. Во-первых, это постоянно действующие факторы, которые оказывают обычно определяющее влияние, и именно они формируют в рядах динамики *основную тенденцию* развития, так называемый *тренд* ***Тр***. Во-вторых, воз-

действие других факторов может проявляться лишь *периодически*. Это вызывает повторяемые во времени *колебания* уровней динамического ряда – либо *сезонного* характера (*сезонный* компонент ***Сз***), либо в виде каких-то *циклических* событий (*циклический* компонент ***Цк***).

Случайная составляющая ***Сл*** отражает действие *разовых* (спародических) факторов, которые проявляются в виде непредсказуемых и нерегулярных изменений уровней рядов динамики.

На рис.6.2 дано структурное изображение основных составляющих динамических рядов.

Таким образом, анализ рядов динамики фактически сводится к оцениванию четырех базовых компонентов помесячных (или поквартальных) временных рядов:

– долгосрочного тренда (тенденции) ***Тр***;

– сезонных колебаний (сезонности) ***Сз***;

– циклической вариации ***Цк***;

– случайных колебаний (нерегулярного компонента) ***Сл***.

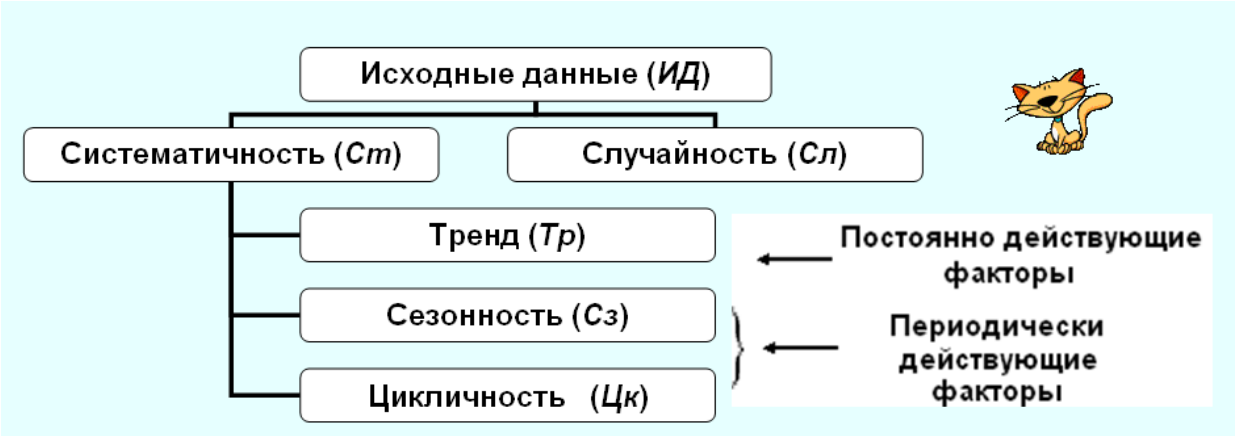


Рис.6.2. Структурная схема базовых компонентов рядов динамики

Поэтому в общем случае базовая модель динамического ряда представляет собой некие числа в этом ряду в виде произведений, полученных путем перемножения указанных компонентов:

$$Исходные\ данные\ (ИД) = Тр \cdot Сз \cdot Цк \cdot Сл$$

Дадим пояснения по поводу этих составляющих:

1. *Долгосрочный тренд* ***Тр*** указывает действительное долгосрочное поведение временно*г*о ряда –, как правило, в виде прямой линии или экспоненциальной кривой. Здесь имеется в виду движение, представляющее нормальное развитие явления (процесса) в течение длительного времени. Это движение является *постоянным* и *медленным*, оно отражает основную тенденцию изменений. Например, возрастание добычи железной руды за последние 50 лет; развитие потребления электроэнергии за последние 10 лет.

2. Точно повторяющийся *сезонный компонент* ***Сз*** определяет влияние *времени года*. Сезонные колебания – это изменения, происходящие в связи с праздниками, различными событиями, обязательными распоряжениями, влияние которых ограничивается определенным сроком. Сезонные изменения бывают порой столь сильными, что нарушают основную линию развития явления. Так, пассажирское движение на Российских железных дорогах очень сильно увеличивается в периоды отпусков (июль-сентябрь). В замерзающих зимой портах в декабре-марте не происходит вообще никакого движения транспорта.

3. Среднесрочный *циклический компонент* ***Цк*** состоит из последовательных повышений и понижений, которые *не* повторяются каждый год. Циклические колебания – это движение по принципу "туда и обратно". В основе лежит последовательная смена состояний подъема и спада, т.е. определенный экономический цикл.

4. Краткосрочный *нерегулярный (случайный) компонент* ***Сл*** представляет остаточную вариацию, которую невозможно объяснить. Это результат случайных колебаний. В нем проявляется действие тех *однократных* событий, которые происходят с течением времени случайно, а не систематически.

Колебания случайного характера выпадают из ритма изменений. Примером могут служить изменения, вызванные последствиями забастовок, финансового краха, издания новых законов в области налогообложения и проч.

В качестве графического пояснения дадим изображение динамического ряда с разложением на все составляющие (рис.6.3).

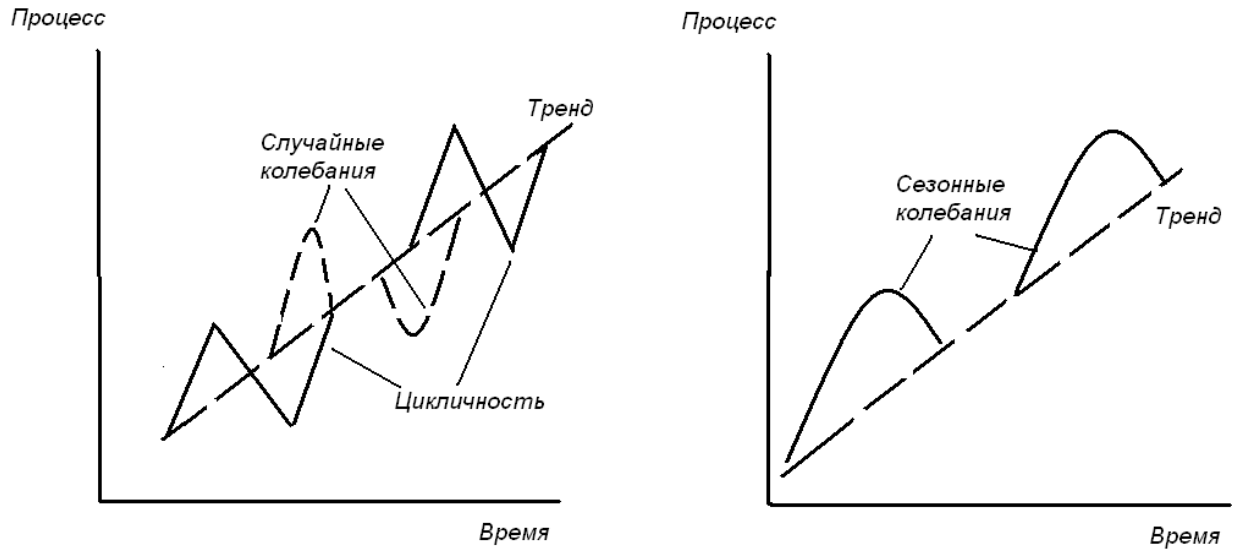


Рис.6.3. Базовые компоненты динамического ряда

Разграничение указанных четырех базовых компонентов (причин хронологических изменений) не всегда удается четко провести.

Итак, мы теперь знаем, что временно*е* развитие процесса в общем случае складывается из нескольких составляющих (их четыре). Нас главным образом интересует *тренд*, поскольку именно он позволяет судить *о динамике* развития изучаемого процесса и дает возможность *заглянуть в будущее*. Однако вокруг него "толпятся" другие факторы, которые путают общую картину и роль тренда может оказаться менее яркой, более размазанной. Вот почему важно уметь выделить:

а) влияние каждого из обсуждаемых факторов;

б) отметить их "весовой" вклад;

в) оценить, наконец, в чистом виде роль самого главного для нас параметра – тренда.

Эти четыре базовых компонента временно*г*о ряда (тренд, сезонность, циклический и нерегулярный компоненты) можно оценивать различными способами. Наиболее удобным и часто применяемым является метод, который называется *отношением к скользящему среднему*.

1. *Скользящее среднее* используется для *устранения сезонных* эффектов путем усреднения по всему году, для *уменьшения нерегулярной* составляющей и получения *комбинации тренда* и *циклического* компонента.

2. Деление *исходного* ряда на *сглаженный* ряд скользящего среднего дает нам *отношение к скользящему среднему*, которое включает как *сезонные*, так и *нерегулярные* значения. Выполняя группирование по времени года,

а затем усреднение в полученных группах, находим *сезонный индекс* для каждого времени года. Исполняя после этого деление каждого значения ряда на соответствующий сезонный индекс для соответствующего времени года, находим значения *с сезонной поправкой*.

3. Регрессия ряда с сезонной поправкой по времени служит для оценки *долгосрочного тренда* в виде прямой линии как функции от времени. Этот тренд (тенденция) *не* отражает сезонных колебаний и дает возможность получить прогноз с сезонной поправкой.

4. Прогнозирование можно выполнять с помощью сезонности тренда. Получая из уравнения регрессии прогнозируемые значения (тренд) для будущих периодов времени и затем умножая их на соответствующий сезонный индекс, мы получаем *прогнозы*, которые отражают как *долгосрочную тенденцию*, так и *сезонное поведение*.

Теперь познакомимся с анализом динамического ряда на конкретном примере.

Студентка четвертого курса специальности "Коммерция" Маша Хорошевская проходила производственную практику в аналитическом отделе солидной торгово-закупочной компании "Максимус", занимающейся поставками металлопроката для предприятий строительного комплекса. С учетом имеющегося спроса на жилищное строительство руководство фирмы заинтересовалось возможными перспективами на предстоящий год. С этой целью аналитическому отделу было поручено спрогнозировать объемы потребления товарной продукции фирмы. Поскольку Маша в рамках изучения университетского курса "Статистика" была знакома с особенностями анализа динамических рядов, то шеф отдела решил поручить такую исследовательскую работу будущему коммерсанту. Смышленая студентка, желая поддержать реноме родного университета и утвердить собственные амбициозные планы, активно включилась в увлекательный творческий процесс…

В табл.6.1 приведены статистические данные о квартальных продажах сортового проката (в млн руб.) за три последних года (2011-2013). По этим данным нужно получить прогнозные соображения относительно перспектив на ближайший год (2014).

Т а б л и ц а 6.1

Исходные данные о продажах сортопрокатной продукции

Год	2011				2012				2013			
Квартал	I	II	III	IV	I	II	III	IV	I	II	III	IV
Продажа, млн руб.	131,6	131, 4	127,1	118,4	149,9	203,8	202,3	196,2	225,7	243,5	265,7	257,3

Вот такая задача поставлена перед нашей студенткой. Постараемся вместе с ней разобраться с этой проблемой.

Итак, приступим…

6.3. Общее описание динамического процесса

Решение сформулированной задачи следует начать с представления исходных данных в формате Excel. Для этого откроим Лист 1 и в нем организуем таблицу, в которую поместим столбиком показатели фактических продаж. В ячейках A1:C13 отметим заголовки, укажем необходимые временные интервалы ("Год", "Квартал"), а также в колонку с названием "Объем продаж" введем сами статистические данные (рис.6.4) .

	A	B	C
1	**Год**	**Квартал**	**Объем продаж, млн руб.**
2	**2011**	**I**	**130,6**
3		**II**	**131,4**
4		**III**	**127,1**
5		**IV**	**118,4**
6	**2012**	**I**	**149,9**
7		**II**	**203,8**
8		**III**	**202,3**
9		**IV**	**196,2**
10	**2013**	**I**	**225,7**
11		**II**	**243,5**
12		**III**	**265,7**
13		**IV**	**257,3**

Рис.6.4. Лист Excel с исходными данными

Построим теперь в графической форме анализируемую зависимость. В этом случае нам будет помогать ***Мастер диаграмм***. Он запускается либо нажатием клавиши на стандартной панели инструментов, либо через команды ***Вставка/Диаграмма*** в строке меню.

Поступим следующим образом:

1. Выделим последнюю колонку, где указаны наши исходные данные – показатели функции (***Объем продаж***).

2. Запустим затем ***Мастер диаграмм*** и выполним рекомендации первого шага – выберем тип диаграммы. В появившемся окне, в левой его части, высветим тип диаграммы – ***График***. Здесь же, нажав кнопку,

Просмотр результата, можно будет посмотреть, как станут выглядеть наши данные на диаграмме выбранного типа.

3. Нажмем на клавишу ***Далее*** и перейдем, следовательно, ко второму шагу. В окне будет активизирована вкладка ***Диапазон данных***. Теперь в кнопке ***Ряды в*** следует указать, что наши данные представлены в ***Столбцах***.

4. В пределах окна второго шага высветим вкладку ***Ряд*** и в строке ***Подписи оси X*** поставим маркер. После этого свернем это окно. Для этого нажмем клавишу справа от поля ввода. В результате можно будет увидеть ту колонку таблицы, где ″сидят″ наши данные по временному диапазону – это столбцы первый (***Год***) и второй (***Квартал***). Выделим оба столбца (без заголовков), после этого вновь нажмем на клавишу – окно полностью раскроется, а на графике по оси абсцисс появятся фактические значения аргумента.

5. Совершим затем следующий, третий шаг (клавиша ***Далее***). Он позволит указать конкретные параметры диаграммы. Запустив вкладку ***Заголовки,*** укажем наименования осей координат (запишем "Время" по оси ***x*** и "Объем продаж, млн руб". по оси ***y***). По желанию можно "украсить" график – добавить или убрать сетку (вкладка ***Линии сетки***), дать необходимые комментарии к графику (вкладка ***Легенда***).

6. Последний шаг – укажем, где желательно поместить график. Для этого вновь нажмем на кнопку ***Далее*** и отметим место расположения его – на имеющемся листе или же отдельном. После завершения этой процедуры последняя приятная операция – прикоснуться к кнопке ***Готово***. Получим график, имеющий вид, представленный на рис.6.5.

Может оказаться, что габариты графика нас решительно не устраивают. Для придания ему более благообразного и удобного вида выделим ***Область диаграммы*** (должны появиться по периметру маркеры-засечки) и поменяем размеры (указатель мыши подведем к маркерам – должны возникнуть двойные стрелки, которые и нужно перемещать). Схожим образом можно изменить габариты самого графика (в пределах имеющейся области диаграммы), выделив ***Область построения диаграммы.***

Теперь внимательно посмотрим на нашу экспериментальную зависимость. На основании визуального анализа можно отметить следующие особенности:

– фиксируются очевидные *сезонные колебания* (наблюдается спад в четвертом квартале);

– просматривается определенная *долговременная тенденция*, а именно: общее повышение объема продаж (кривая ползет вверх);

– наблюдается некоторая *нерегулярность поведения*.

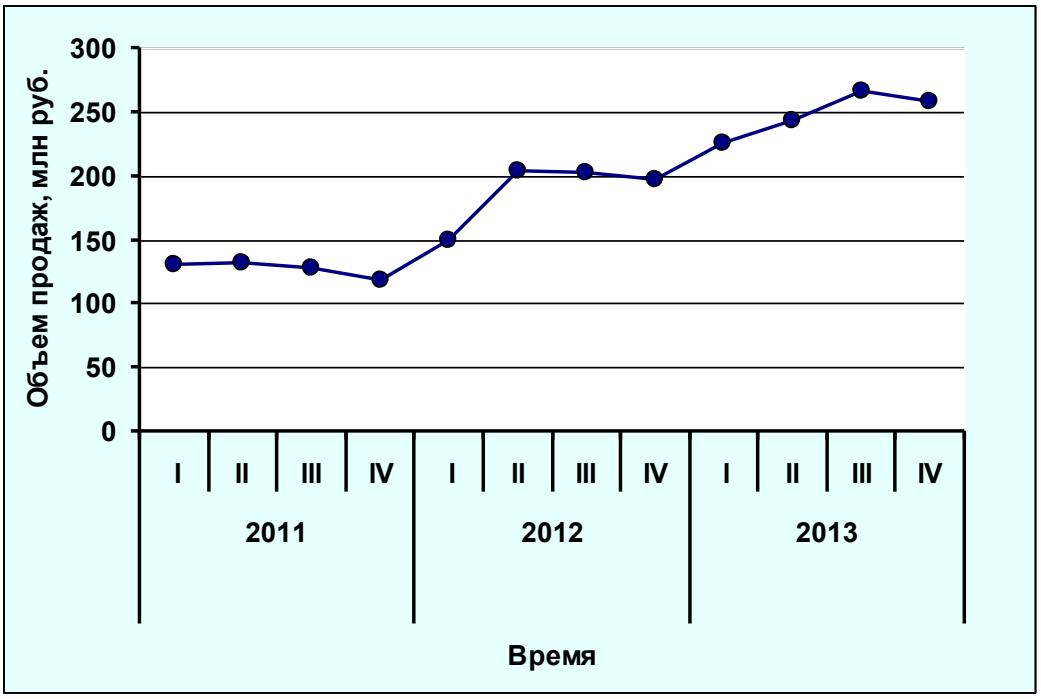


Рис.6.5. График динамического ряда поквартальных продаж

7. Нанесем линию тренда, для этого подведем стрелку мыши к линии графика и щелкнем правой клавишей. Появится окно ***Формат рядов данных.*** Выделим опцию ***Добавить линию тренда***, в результате появится всплывающее окно ***Линия тренда***. На вкладке ***Тип*** выберем похожий на нашу кривую график-шаблон. Для данного случая вполне подходящей оказывается линейная зависимость. Перейдем затем к вкладке ***Параметры*** и укажем засечками команды ***Показать уравнение на диаграмме*** и ***Поместить на диаграмме величину достоверной аппроксимации*** $\boldsymbol{R^2}$. После нажатия клавиши ***ОК*** график примет окончательный вид (рис.6.6). Отметим, что наша экспериментальная кривая характеризуется довольно большим показателем статистического соответствия с линией тренда – аппроксимирующий коэффициент (коэффициент детерминации) $\boldsymbol{R^2}$ составляет 0,90. Это свидетельствует

о существовании сильной корреляционной связи между изучаемыми совокупностями.

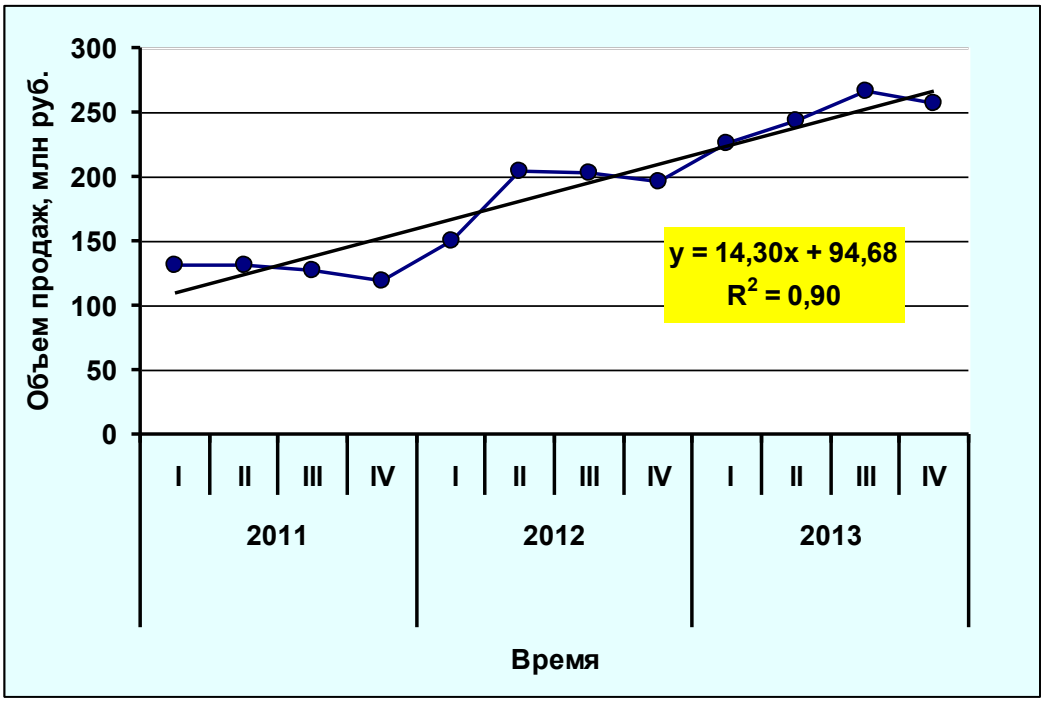


Рис.6.6. Исходные данные и линия тренда с уравнением регрессии

6.4. Вычисление скользящего среднего

Наша цель состоит в том, чтобы выделить четыре базовых компонента ряда динамики. Разложение исходного динамического ряда на эти составляющие и позволяет получить четкую картину влияния каждого компонента.

Начнем с *усреднения* данных за год, чтобы а) *избавиться* от *сезонного* компонента и б) *уменьшить случайный (нерегулярный)* компонент.

Скользящее среднее представляет собой *новый* ряд, полученный путем усреднения соседних наблюдений динамического ряда и перехода к следующему периоду времени – в результате получится более гладкий ряд. Выполняя усреднение данных за целый год, мы приходим к тому, что *вклад сезонных* колебаний – независимо от времени года – остается практически *одинаковым*.

Суть метода скользящего среднего – замена *абсолютных* данных *средними арифметическими* за определенные периоды. Расчет средних величин ведется способом *скольжения*, т.е. постепенным исключением из принятого

периода скольжения первого уровня и включением следующего. Здесь сглаживание динамического ряда можно осуществить, например, методом трехчленной (скажем, за три месяца, т.е. за квартал) или четырехчленной скользящей средней.

Для метода скользящего среднего условно можно записать следующие процедуры:

– выполнить усреднение соседних наблюдений за определенный период (этот временной интервал принято называть *"окном"*), например, год;

– осуществить операцию скольжения, т.е. обеспечить переход к следующему среднему путем исключения из принятого "окна" первого уровня и включения следующего – получается, что выбранный интервал ("окно") скользит вдоль ряда.

Схематически это показано на рис.6.7.

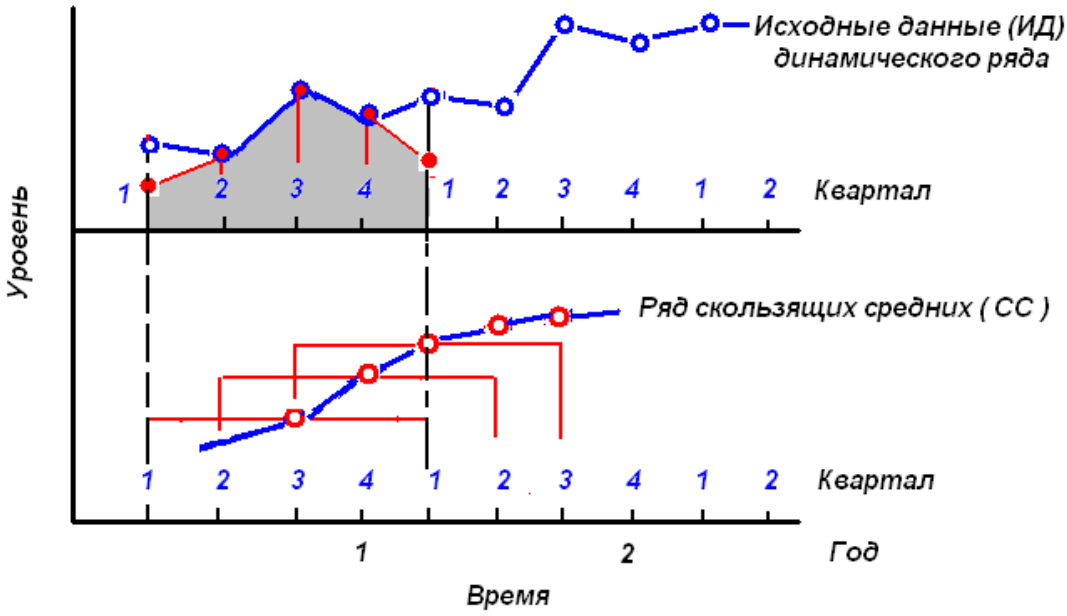


Рис.6.7. Сглаживание динамического ряда методом скользящего среднего

Таким образом, скользящее среднее ***СС*** можно охарактеризовать как показатель, учитывающий влияние тренда ***Тр*** и цикличности ***Цк***:

$$\text{Скользящее среднее (СС)} = Тр \cdot Цк$$

Найти скользящее среднее значение для *поквартальных* данных за определенный период времени можно следующим образом.

1. Начнем с текущего значения y_i и добавим к нему значения его "соседей" справа y_{i+1} и слева y_{i-1}.

2. Прибавим затем *половину* значений следующих "соседей", т.е. получится 0,5 y_{i+2} и 0,5 y_{i-2}.

3. Имеющуюся сумму разделим на 4.

Такое взвешенное среднее нужно для того, чтобы *интервал* по обе стороны от базового периода времени был *симметричным* и вместе с тем охватывал в точности *данные за один год*. Взвешивая крайние точки коэффициентом 0,5, мы гарантируем, что этот квартал учтен в скользящем среднем точно так же, как и другие кварталы.

Следовательно, можно записать так:

$$\bar{y}_i = \frac{0{,}5y_{i-2} + y_{i-1} + y_i + y_{i+1} + 0{,}5y_{i+2}}{4} \text{ и т.д.}$$

Пояснение. В отечественной литературе по статистике рекомендуется вторых (и третьих) "соседей" прибавлять справа и слева целиком (не делить пополам):

$$\bar{y}_1 = \frac{y_1 + y_2 + y_3}{3};\ \bar{y}_2 = \frac{y_2 + y_3 + y_4}{3};\ \ldots \text{ или}$$

$$\bar{y}_1 = \frac{y_1 + y_2 + y_3 + y_4}{4};\ \bar{y}_2 = \frac{y_2 + y_3 + y_4 + y_5}{4} \text{ и т.д.}$$

В этой связи использование метода скользящего среднего учитывает такую особенность, как размер окна сглаживания, длина которого может выражаться как четным, так и нечетным числом. В случае четного числа усредненное значение нельзя приписать какому-то определенному моменту времени, поскольку средняя величина может быть отнесена только к середине между двумя смежными датами, находящимися в середине окна сглаживания. Для определения сглаженных уровней при размере окна в виде четного числа применяется так называемых *метод центрирования*, который заключается в нахождении среднего арифметического из двух смежных скользящих средних для отнесения полученного уровня к определенной дате.

А теперь приступим собственно к решению нашей задачи. Напомним, что нам нужно попытаться разложить временной ряд на его составляющие.

Расчет выполняется для случая суммирования "без крайних половинок".

Вначале займемся сглаживанием динамического ряда с помощью метода скользящего среднего. Предпримем следующие шаги.

1. Введем данные, приведенные в столбцах A:C (рис.6.4), на новый лист. Для этого их скопируем и перенесем в Лист 2. Зарезервируем в колонках A и B ячейки для четырех кварталов 2014 года (это в дальнейшем нам потребуется на этапе построения прогноза).
2. Отметим заголовки (метки) ***Раньше_СС***, ***Позже_СС*** и ***Центрированное_СС*** в ячейки D1:F1, как показано на рис.6.8.
3. Выделим ячейку D4 и введем формулу **=СРЗНАЧ(C2:C5)**. Указанное среднее первых четырех кварталов соответствует точке между вторым и третьим кварталами. Оно расположено в строке третьей четверти и будет обозначаться как ***Раньше_СС***.
4. Затем выделим ячейку E4 и введем формулу **=СРЗНАЧ(C3:C6)**. Тем самым будет рассчитано среднее кварталов со второго по пятый и этому станет соответствовать точка между третьим и четвертым кварталами. Это среднее располагается также в строке третьей четверти и примет обозначение ***Позже_СС***.
5. Выделим ячейку F4 и введем формулу **=СРЗНАЧ(D4:E4)**. Будет получено среднее значений ***Раньше_СС*** и ***Позже_СС***, которое покажет центрированное значение для третьего квартала.
6. Теперь выделим ячейки D4:E4 и щелкнем по маркеру заполнения в правом нижнем углу выделенной области и протянем его к ячейке F11. Полученные результаты представим с одним разрядом после запятой (рис.6.8).

	A	B	C	D	E	F
1	**Год**	**Квартал**	**Объем продаж, млн руб.**	**Раньше_СС**	**Позже_СС**	**Цен_СС**
2		I	130,6			
3	2011	II	131,4			
4		III	127,1	126,9	131,7	129,3
5		IV	118,4	131,7	149,8	140,8
6		I	149,9	149,8	168,6	159,2
7	2012	II	203,8	168,6	188,1	178,3
8		III	202,3	188,1	207,0	197,5
9		IV	196,2	207,0	216,9	212,0
10		I	225,7	216,9	232,8	224,9
11	2013	II	243,5	232,8	248,1	240,4
12		III	265,7			
13		IV	257,3			
14		I				
15	2014	II				
16		III				
17		IV				

Рис.6.8. Лист Excel с центрированными скользящими средними

Чтобы на самой диаграмме отобразить скользящее среднее, проделаем следующее.

7. Выделим ячейки C1:C13. Удерживая нажатой клавишу Ctrl, активизируем диапазон ячеек F1:F13 и затем запускаем ***Мастер диаграмм***. На шаге 1 ***Мастер диаграмм*** на вкладке ***Стандартные*** укажем ***Тип График*** и ***Вид График*** с маркерами, помечающими экспериментальные точки. Перейдем к следующему шагу 2 (***Источник данных диаграммы***). На вкладке ***Ряд*** щелкнем в строке ***Подписи оси X*** и перетащим диапазон ячеек A2:B13. На шаге 3 (***Параметры диаграммы***) для вкладки ***Заголовок*** введем названия координатных осей. После этого – клавиша ***Готово***. Полученный результат можно видеть на рис.6.9.

Приведенный график следует толковать следующим образом: удалось устранить сезонные и случайные колебания объемов продаж сортовой металлопродукции, однако остался тренд и сохранилось влияние циклического компонента.

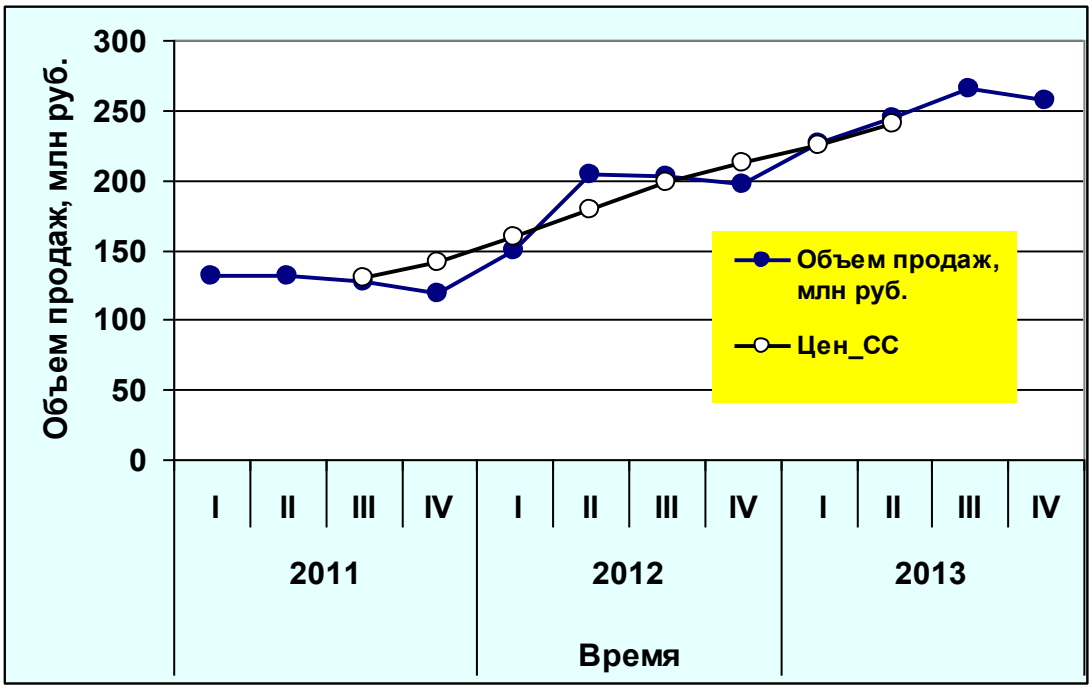


Рис.6.9. График фактических продаж и центрированных скользящих средних

Чтобы выделить *сезонное* поведение, прежде всего, нужно получить отношение исходных значений к скользящему среднему. Именно отсюда происходит название "отношение к скользящему среднему". Полученный ре-

зультат будет *включать сезонный* и *случайный* компоненты, поскольку скользящее среднее *исключает* из данных *тренд* и *циклическую* составляющую.

Сказанное можно представить в такой записи:

$$\text{Отношение к скользящему среднему} = \frac{ИД}{СС} = \frac{Тр \cdot Сз \cdot Цк \cdot Сл}{Тр \cdot Цк} = Сз \cdot Сл$$

8. Введем метки ***Отношение***, ***СрОтношение*** и ***Нормированные*** в ячейки G1:I1. Выделим ячейку G4 и запишем формулу **=C4/F4**. При выделенной ячейке G4, нажмем клавишу Enter, затем щелкнем по маркеру заполнения и перетащим к ячейке G15. Результаты этой манипуляции предстанут в столбце G рис.6.10.

Данные числа являются отношением фактических продаж (абсолютных данных) к скользящему среднему. Например, число 0,841 в ячейке G5 показывает, что фактические продажи за четвертый квартал 2011 года составили 84,1% от средних продаж в течение года.

	A	B	C	D	E	F	G	H	I
1	Год	Квартал	Объем продаж, млн руб.	Раньше_СС	Позже_СС	Цен_СС	Отнош	СрОтнош	Нормир
2		I	130,6					0,973	0,988
3	2011	II	131,4					1,078	1,095
4		III	127,1	126,9	131,7	129,3	0,983	1,004	1,020
5		IV	118,4	131,7	149,8	140,8	0,841	0,883	0,897
6		I	149,9	149,8	168,6	159,2	0,942	3,938	4,000
7	2012	II	203,8	168,6	188,1	178,3	1,143		
8		III	202,3	188,1	207,0	197,5	1,024		
9		IV	196,2	207,0	216,9	212,0	0,926		
10		I	225,7	216,9	232,8	224,9	1,004		
11	2013	II	243,5	232,8	248,1	240,4	1,013		
12		III	265,7						
13		IV	257,3						
14		I							
15	2014	II							
16		III							
17		IV							
18									

Рис.6.10. Лист Excel с сезонными индексами

6.5. Анализ сезонных колебаний

Теперь, чтобы устранить *случайный (нерегулярный)* компонент, мы усредним эти значения для каждого сезона. *Сезонный* компонент проявляет-

ся, поскольку он присутствует ежегодно, тогда как нерегулярный компонент, как правило, удается усреднить.

Нужно будет рассчитать так называемый сезонный индекс, который представляет собой усредненную сезонную компоненту на весь рассматриваемый период времени (для нашего примера – это три года). Для этого необходимо выбрать все отношения скользящего среднего за конкретный период, например, третий квартал, их просуммировать и затем разделить на общее число этих кварталов за рассматриваемый период (их будет два). И так следует поступить с остальными временны*ы*ми интервалами.

В удобном виде это можно представить так:

$$\text{Сезонный индекс (СИн)} = \frac{\text{Сумма } \frac{\text{ИД}}{\text{СС}} \text{ за соответствующий период}}{\text{Общее число } n \text{ этого периода}}$$

9. Выделим ячейку H2 и введем **= (G6+G10)**/2, а затем при выделенной этой ячейке щелкнем по маркеру заполнения и протянем его к ячейке H3.

10. Активизируем теперь уже ячейку H4 и введем формулу **=(G4+G8)/2**, выделенную ячейку (после клавишы Enter) с помощью маркера перетащим в позицию H5. Результат можно увидеть в столбце H (рис.6.10). Здесь даны итоговые значения ***Отношения*** квартала для всех лет.

11. Теперь выделим ячейку H6 и запустим опцию ***Автосумма*** на инструментальной панели (значок ∑). При отсутствии сезонного компонента индекс должен быть равен 1,00, поэтому сумма всех четырех индексов должна составлять 4. Для нормирования средних отношений (чтобы их сумма равнялась четырем) выделим ячейку I2 и введем формулу **=H2*4/$H$6**. При выделенной I2 щелкнем по маркеру заполнения и протянем его к ячейке I5.

Выделим ячейку I6 и щелкнем дважды по инструменту ***Автосумма***. Как видно, сумма сезонных индексов в столбце I будет равна 4 (рис.6.10).

Пояснение. Например, если рассмотреть третьи кварталы соответственно 2011 и 2012 годов, то для них расчет будет выглядеть так: (0,983+1,024)/2=1,004. После нормирования этот показатель станет равным 1,020. Это и есть сезонный индекс

для третьего квартала. Он был получен путем усреднения отношений за третий квартал по всем рассматриваемым годам.

Схожим образом выполнятся вычисления сезонного индекса и для других кварталов.

12. Построим график в виде столбиковой диаграммы, иллюстрирующий типичную картину изменения сезонных индексов в течение года (рис.6.11). Прием уже знакомый – выделим ячейки H2:H5, запустим ***Мастер диаграмм***, а далее уже привычным способом. Заметим, что при выполнении первого шага выберем график в виде гистограммы, затем при втором шаге в окне ***Подписи оси X*** отметим диапазон B2:B5 (там указаны номера кварталов). И еще одно замечание. Для удобства столбики дополним числовыми значениями индексов. Поступим так: при выполнении третьего шага выберем вкладку ***Подписи данных*** и активизируем окно ***Значения***.

Можно затем придать более приятный вид диаграмме – изменить ширину столбцов, масштаб по оси ординат. Для этого последовательно следует активизировать опции ***Формат рядов данных*** и ***Формат оси***. Чтобы вызвать эти команды, нужно проделать следующее. В первом случае подведем маркер к какому-либо столбцу и, нажав правую клавишу, вызовем нужное контекстное меню. Во втором случае проделаем аналогичную процедуру, только маркером следует предварительно указать ось ординат.

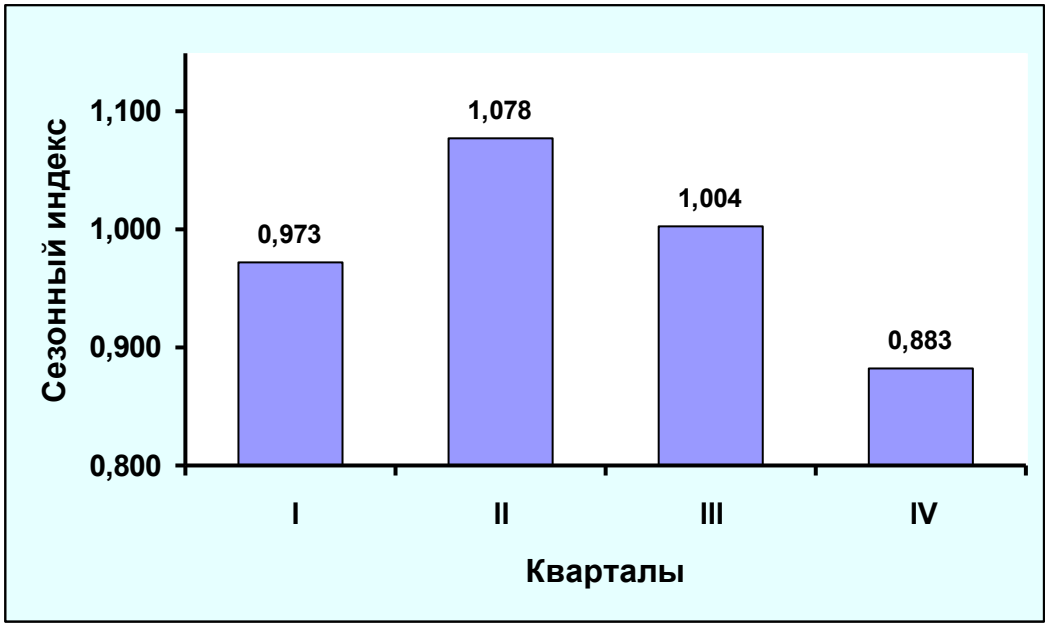


Рис.6.11. Поквартальное изменение сезонных индексов

После того как вычислен каждый сезонный индекс, его можно использовать везде – даже там, где нельзя вычислить скользящее среднее, поскольку, по определению, сезонные колебания в точности повторяются каждый год.

Рис.6.11 иллюстрирует типичную картину сезонных колебаний в течение года. Представленный график надлежит понимать следующим образом. Сезонные индексы для рассматриваемой ситуации показывают, что объемы продаж металлопроката, как правило, достигают пика во втором квартале (на 7,8 % выше среднегодового показателя, или так называемого типичного квартала). Затем они падают до минимума в четвертом квартале (на 11,7 % ниже уровня типичного квартала), а затем снова повышаются вплоть до следующего второго квартала. И такая картина повторяется из года в год для данного исследуемого процесса.

6.6. Поправка на сезонный фактор

Поправка на сезонные колебания – она *устраняет* из результатов измерения ожидаемый *сезонный компонент*. Это позволяет сравнивать один квартал или месяц с другим (после внесения поправки на сезон), выявляя тем самым те или иные скрытые тенденции.

Поясним сказанное следующим примером. Так, для розничной торговли декабрь является обычно наиболее благополучным месяцем. Если объём продаж в декабре оказывается выше по сравнению с ноябрем, то это вполне *ожидаемый* результат. Но если объём продаж в декабре оказывается *выше* даже по сравнению с ожидаемыми показателями, это значит, что даже с учетом поправки на сезонные колебания продажи *существенно возросли*. Если же объём продаж в декабре оказался выше, чем ноябре, но все же *меньше* ожидаемого, то можно говорить, что с поправкой на сезонные колебания декабрьские продажи на самом деле *снижаются*.

Чтобы найти некоторое значение с поправкой на сезонные колебания, достаточно разделить исходные данные на сезонный индекс для соответствующего месяца или квартала.

$$\text{Значение с поправкой на сезон} = \frac{\text{Исходные данные (ИД)}}{\text{Сезонный индекс}} = \frac{Тр \cdot Сз \cdot Цк \cdot Сл}{СИн} = Тр \cdot Цк \cdot Сл$$

Продолжим наши расчеты.

13. Укажем заголовки ***СзИндекс***, ***Тренд, Периоды*** и ***Прогноз*** в ячейках J1:M1. Затем выделим ячейки I2:I5 и нажмем кнопку ***Копировать*** (или по-другому: щелкнем правой клавишей мышки и в появившемся контекстном меню выберем опцию ***Копировать***). Выберем ячейку J2, щелкнем правой клавишей и укажем ***Специальная вставка*** в контекстном меню. В диалоговом окне ***Специальная вставка*** отметим ***Вставить значения*** и ***Нет*** в разделе ***Операция***. Пункты ***Пропускать пустые ячейки*** и ***Транспортировать*** оставим выключенными. После чего – клавиша ***ОК***.

14. Скопируем содержимое ячеек J2:J5 и вставим их в ячейки J6, J10 и J14. Получим столбец J, в котором периодически повторяются четыре числа – сезонные индексы (коэффициенты сезонности).

15. Выделим теперь ячейку K2 и введем формулу **=C2/J2**. При активизированной ячейке K2 щелкнем маркером заполнения и протянем его до позиции K13. Теперь в диапазоне K2:K13 будут находиться сезонно-скорректированные данные.

16. Отметим ячейки K2:K13 и опцию ***Копировать,*** затем щелкнем правой клавишей и в контекстном меню повторим знакомую процедуру: ***Специальная вставка/Вставить значения/Операция/Нет***. Проигнорируем пункты ***Пустые ячейки*** и ***Транспортировать*** и затем ***ОК***.

17. При выделенных ячейках K2:K13 активизируем маркер заполнения и протащим его к ячейке K17. Результаты будут представлены в столбце K. В этом случае Excel дополнит ряд чисел K2:K17, используя линейный тренд (рис.6.12).

Прокомментируем выполненные процедуры на конкретном примере. Если обратиться к таблице на рис.6.12, то видно, что фактический объем продаж во втором квартале 2013 года составил 243,5 млн руб. (см. ячейку C11), а сезонный индекс для этого же периода равнялся 1,095 (ячейка J11). Результат деления первого числа на второе, равный 222,4 млн руб. (ячейка K11), составит объем продаж с поправкой на сезонные колебания.

C	D	E	F	G	H	I	J	K	L	M
Объем продаж, млн руб.	Раньше_СС	Позже_СС	Цен_СС	Отнош	СрОтнош	Нормир	СезИндекс	Тренд	Периоды	Прогноз
130,6					0,973	0,988	0,988	132,2	1	
131,4					1,078	1,095	1,095	120,0	2	
127,1	126,9	131,7	129,3	0,983	1,004	1,020	1,020	124,7	3	
118,4	131,7	149,8	140,8	0,841	0,883	0,897	0,897	131,9	4	
149,9	149,8	168,6	159,2	0,942	3,938	4,000	0,988	151,7	5	
203,8	168,6	188,1	178,3	1,143			1,095	186,1	6	
202,3	188,1	207,0	197,5	1,024			1,020	198,4	7	
196,2	207,0	216,9	212,0	0,926			0,897	218,6	8	
225,7	216,9	232,8	224,9	1,004			0,988	228,4	9	
243,5	232,8	248,1	240,4	1,013			1,095	222,4	10	
265,7							1,020	260,6	11	
257,3							0,897	286,7	12	
							0,988	287,2	13	283,8
							1,095	302,4	14	331,1
							1,020	317,6	15	323,8
							0,897	332,8	16	298,7

Рис.6.12. Лист Excel с сезонными индексами и прогнозом

Как видно, результат с поправкой на сезон оказался меньше фактического объёма продаж. Дело в том, что объём продажи во втором квартале, как правило, выше по сравнению с типичным кварталом года. В сущности, мы заранее можем рассчитывать на то, что объём продаж во втором квартале будет примерно на 9,5% выше (исходя из сезонного индекса, равного 1,095). Деление на сезонный индекс нивелирует влияние этой ожидаемой сезонной флуктуации. В результате объём продажи во втором квартале приводится в соответствие с типичным кварталом года (т.е. снижая его).

В следующем квартале (третьем, 2013 год) объём продажи с поправкой на сезонные колебания равняется 265,7/1,020=260,7. Видно, что наблюдается повышение фактического объёма продаж (с 243,5 во втором квартале до 265,7 в третьем, т.е. на 9,2%). Если же воспользоваться поправкой на сезон, то окажется, что объем продажи и в этом случае возрос, но более существенно – с 222,4 до 260,6, что составит 17,3%. Это говорит о том, что отмеченное нами повышение объема продаж на самом деле оказалось более серьезным, чем можно было ожидать для этого времени года.

Еще пример. Так, бросается в глаза значительное снижение объема продаж в четвертом квартале 2013 года (с 265,7 упало до 257,3, что составит –3,2 %). Но если воспользоваться поправкой на сезон, то оказывается, что в этом квартале фиксируется вполне приличный рост (с 260,7 до 286,7, т.е. на 7,7 %!).

Введение сезонной поправки, таким образом, позволяет получить более объективное представление о реальном поведении исследуемой зависимости. В нашем случае она показывает, что мы имеем дело с "настоящим" ростом объёма продаж, а не просто с сезонным увеличением.

18. Теперь нужно отобразить фактические значения продаж, сезонно-скорректированные продажи, а также линейную экстраполяцию на диаграмме. Для этого выделим ячейки C1:C17 и, удерживая клавишу Ctrl, отметим ячейки K2:K17. Запустим ***Мастера диаграмм*** и получим график, который показан на рис.6.13.

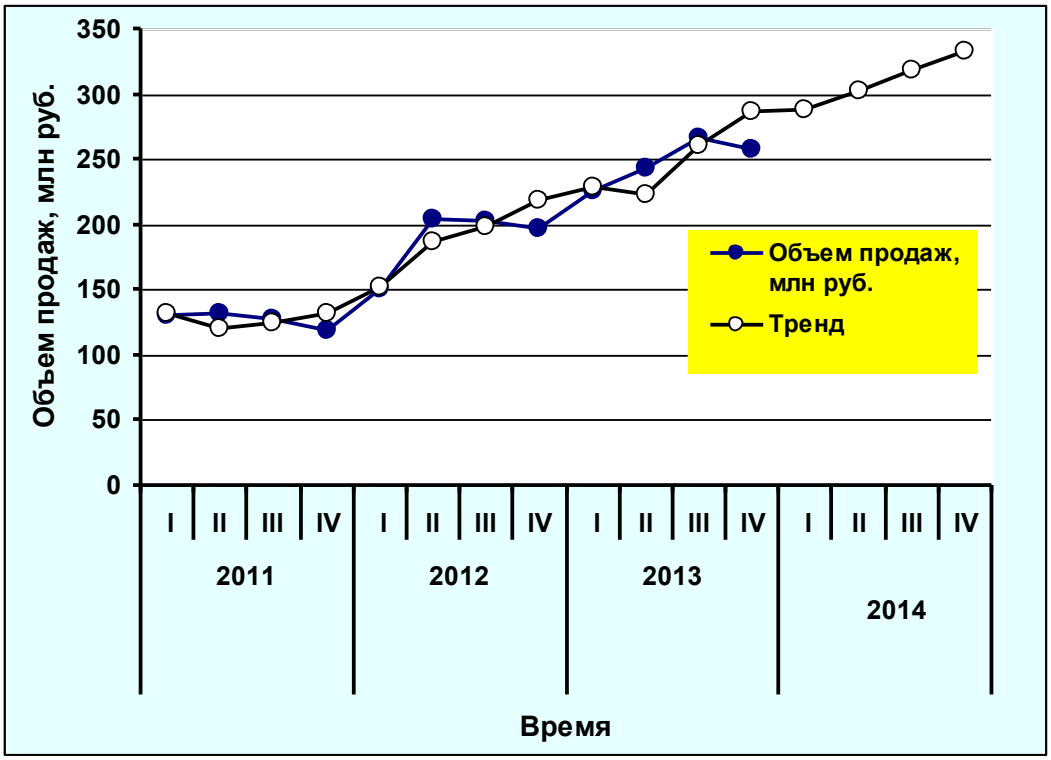


Рис.6.13. Экстраполяция сезонно-скорректированных значений продаж

Графическое представление объемов продаж с поправкой на сезонные колебания показывает, что динамический ряд оказывается более гладким, чем исходные данные, поскольку нам удалось избавиться от сезонных отклонений.

Итак, только сейчас, на этом этапе, удалось полностью "очистить" наши исходные данные от сезонного поведения. Вместе с тем сохраняется маскирующее воздействие на тренд других составляющих – цикличности и нерегулярности.

Продолжим анализ.

6.7. Долгосрочный тренд и прогноз с поправкой на сезонность

Когда динамический ряд демонстрирует долгосрочную линейную тенденцию к нарастанию или снижению, для оценки этой тенденции и прогнозирования будущего можно воспользоваться регрессионным анализом.

Регрессионный анализ в этом случае сводится к следующему. Для прогнозирования ряда, в котором учитывается поправка на сезонность (переменная ***y***), используется период времени (переменная ***x***). Результирующее уравнение регрессии будет представлять долгосрочный тренд. Подставляя будущие временны*е* периоды в качестве новых значений ***x***, мы получим возможность экстраполировать эту долгосрочную тенденцию в будущее.

При описании временны*х* рядов важно выбрать числа так, чтобы они были распределены равномерно. Этого можно добиться, если воспользоваться числами 1, 2, 3,… для представления непосредственно *в виде номера* временно*го* периода (квартала или месяца).

Поэтому в нашей таблице (рис.6.12) используем зарезервированную колонку ***Период***, где укажем в виде номеров (от 1 до 16) кварталы за весь анализируемый временно*й* интервал, т.е. 2011-2014 годы.

Построим график. Для этого воспользуемся уже знакомыми командами (см. пункты 1-7). Используем данные, расположенные в столбцах C, K и L. Фактически получится график (рис.6.14), аналогичный показанному ранее на рис.6.5. Отличие только в том, что в данном случае уравнение регрессии построено на основании данных, учитывающих сезонные колебания и рассматривающих более продолжительный временно*й* интервал (четыре года, а не три). Зависимость, как и следовало ожидать, хорошо описывается линейной функцией и характеризуется более высокой достоверностью (коэффициент детерминации $\boldsymbol{R^2}$ близок к 1).

Рис.6.14. Тренд с учетом сезонности и линия регрессии

Небольшое добавление. Чтобы на полях графика разместить поясняющие надписи ("Ряд с поправкой на сезонность", "Линия регрессии"), нужно запустить опцию ***Исходные данные*** во вкладке ***Ряд***, и заполнить соответствующей фразой позицию ***Имя.*** В левом окошке ***Ряд*** появится данная запись. При необходимости можно вводить следующую фразу, активизировав функцию ***Добавить***.

Таким образом, уравнение регрессии, построенное посредством метода наименьших квадратов, имеет следующий вид:

$$y = 89{,}7 + 15{,}2x$$

Оно показывает, что объемы продаж сортовой металлопродукции компании "Максимус" увеличиваются в среднем на 15,2 млн руб. за квартал.

Этот долгосрочный тренд легко прогнозировать, подставляя в уравнение регрессии соответствующий временно́й период. Например, чтобы найти значение тренда для первого квартала 2014 года, нужно использовать значение x=13, которое будет представлять период времени, следующий за окончанием нашего временно́го ряда. В этом случае прогноз будет иметь следующий вид:

$$y = 89{,}7 + 15{,}2 \cdot 13 = 287{,}2 \text{ (в млн руб.)}.$$

В нашей основной таблице (рис.6.12) представлены прогнозируемые значения (показатели долгосрочного тренда и его прогноз на один год вперед по отношению к имеющимся у нас данным).

Таким образом, линия тренда отражает поведение динамического ряда: с одной стороны, учитывается поправка на сезонные колебания, а с другой, – благодаря экстраполяции – определяется прогноз на будущее (с поправкой на сезонность).

6.8. Прогноз: тренд с учетом сезонности

Чтобы в *полной мере* иметь возможность *прогнозировать будущее*, нужно учесть *сезонность* в *долгосрочном тренде*. Иначе говоря, следует *вернуть* ему *ожидаемую сезонную изменчивость*. Для этого достаточно умножить значение тренда на значение сезонного индекса для того периода времени, который подлежит прогнозу. Фактически этот процесс является обратным по отношению к внесению поправки на сезонные колебания.

Результирующий прогноз включает долгосрочный тренд и сезонную вариацию:

$$Прогноз = Тр \cdot СИн$$

Чтобы предсказать объёмы продаж компании "Максимус" за первый квартал 2014 года, достаточно умножить значение тренда, равное 287,2 (вычисляется с помощью уравнения регрессии для 13-го временн*о*го периода), на сезонный индекс для первого квартала, равный 0,988:

$$287{,}2 \cdot 0{,}988 = 283{,}8 \text{ (в млн руб.)}.$$

Мы проделали такую рутинную операцию, чтобы было понятно, каким образом получились прогнозные показатели. А теперь проделаем это же самое, используя возможности Excel.

19. Совместим сезонный компонент и тренд в прогнозе. Для этого выделим ячейку M14 и введем формулу **=J14*K14**. Затем при выделенной ячейке M14 дважды щелкнем маркером заполнения – итоговые данные можно видеть на рис.6.12. Тем самым получим прогнозные данные на год вперед (на 2014-й) по отношению к имеющимся данным.

20. Перейдем к заключительному этапу – построим график, иллюстрирующий наши фактические данные по поводу продаж и, самое главное, взгляд в будущее, т.е. долгожданный прогноз. Выделим ячейки C1:C17 и, удерживая нажатую клавишу Ctrl, отметим диапазон L1:L17. Затем запустим ***Мастер диаграмм*** и далее в привычном режиме… Итоговый результат можно видеть на рис.6.15.

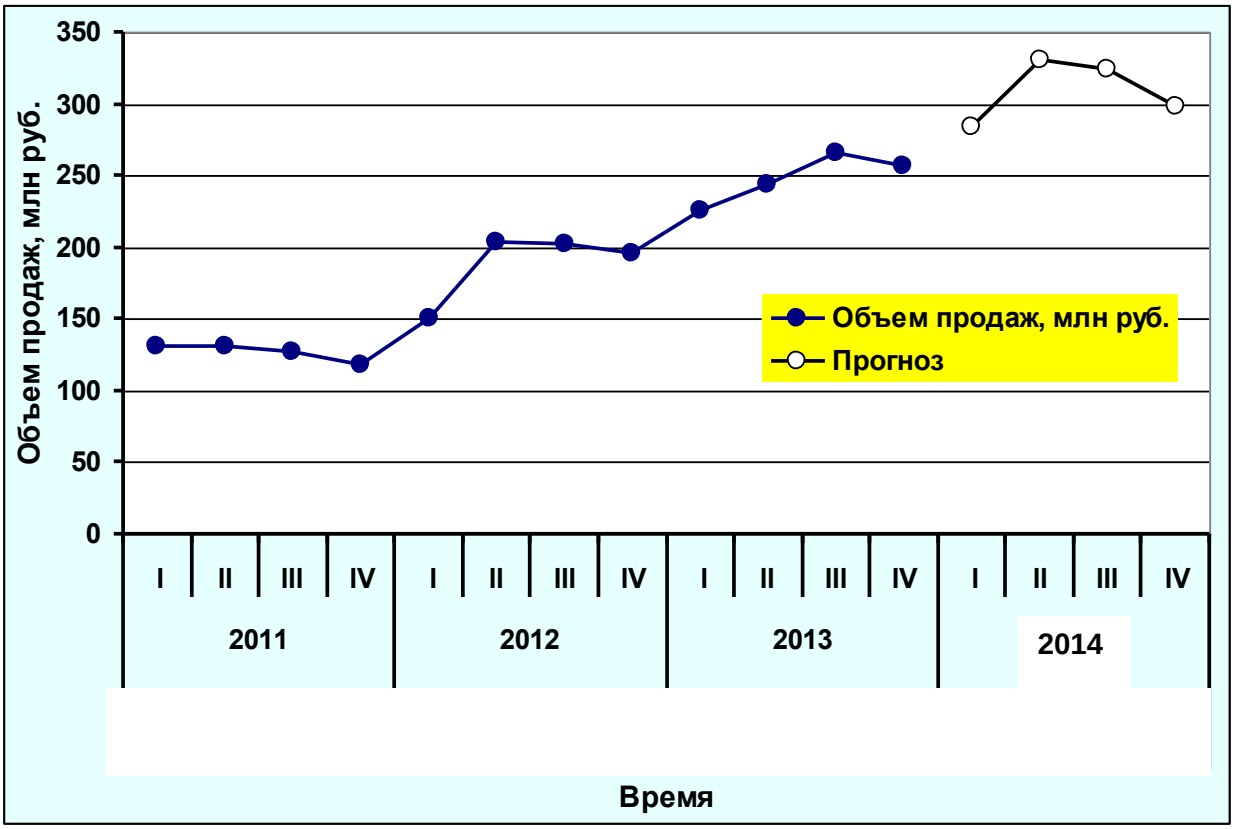


Рис.6.15. Фактические данные объема продаж и результаты прогнозирования

Таким образом, график показывает, как этот тренд, учитывающий сезонность, отражает анализируемый нами ряд и продолжается (путем экстраполяции) вправо, обеспечивая достаточно надежные прогнозы, включающие ожидаемое сезонное падение объёмов продаж.

Итак, на этом наше исследование закончено. Полагаем, что руководство компании надлежащим образом оценило усердие и способности Маши Хорошевской и прозрачно намекнуло на желательность иметь в своем штате такого полезного сотрудника…

И последнее. Напомним, что практически все прогнозы не очень-то достоверны. В конце концов, нерегулярный компонент невозможно предсказать по определению.

Однако положительная роль прогнозов заключается хотя бы в том, что они позволяют выявить долгосрочные тенденции нарастания (или убывания), а также повторяющиеся сезонные колебания. В нашем случае было бы заманчиво провести сравнения между фактическими значениями объемов продаж в 2014 году с тем, что дает прогноз. Тогда можно будет достаточно определенно судить о надежности наших прогнозных предсказаний.

БИБЛИОГРАФИЧЕСКИЙ СПИСОК

1. Сигал Э. Практическая бизнес-статистика. – М.: издательский дом «Вильямс», 2002. – 1056 с.

2. Макарова Н. В., Трофимец В. Я. Статистика в Excel: учебное пособие. – М.: Финансы и статистика, 2002. – 192 с.

3. Мидлтон М. Р. Анализ статистических данных с использованием Microsoft Excel для Office XP. – М.: БИНОМ. Лаборатория знаний, 2005. – 296 с.

4. Нельсон С. Анализ данных в Excel для «чайников». – М.: издательский дом «Вильямс», 2002. – 302 с.

5. Годин А. М. Статистика: Учебник. – М.: Издательско-торговая корпорация «Дашков и К°», 2002. – 368 с.

6. Хайкин Б. Е. Построение аппроксимационных математических моделей в условиях обработки металлов давлением. Учебное пособие.– Свердловск: УПИ, 1991. – 101 с.

ПРИЛОЖЕНИЯ

Статистико-математические таблицы

Приложение 1

Значения коэффициента Стьюдента (*t*-критерия) при уровне значимости α и числе измерений *n*

n \ α	0,1	0,05	0,01
2	6,314	12,706	63,657
3	2,920	4,303	9,925
4	2,353	3,182	5,841
5	2,132	2,776	4,604
6	2,015	2,571	4,032
7	1,943	2,447	3,707
8	1,895	2,365	3,499
9	1,860	2,306	3,355
10	1,833	2,262	3,250
11	1,812	2,228	3,169
12	1,796	2,201	3,106
13	1,782	2,179	3,055
14	1,771	2,160	3,012
15	1,761	2,145	2,977
16	1,753	2,131	2,947
17	1,746	2,120	2,921
18	1,740	2,110	2,898
19	1,734	2,101	2,878
20	1,729	2,093	2,861
21	1,725	2,086	2,845
22	1,721	2,080	2,831
23	1,717	2,074	2,819
25	1,711	2,064	2,797
27	1,706	2,056	2,779
29	1,701	2,048	2,763
31	1,697	2,042	2,750
40	1,684	2,021	2,704
60	1,671	2,000	2,660
120	1,658	1,980	2,617
∞	1,645	1,960	2,576

Приложение 2

Значение эталонного (табличного) показателя $\tau_{табл}$, для оценки выскакивающих значений (отсева грубых промахов) при уровне значимости α и числе измерений *n*

n \ α	0,1	0,05	0,01
3	1,41	1,41	1,41
4	1,65	1,69	1,72
5	1,79	1,87	1,96
6	1,89	2,00	2,13
7	1,97	2,09	2,27
8	2,04	2,17	2,37
9	2,10	2,24	2,46
10	2,15	2,29	2,54
11	2,19	2,34	2,61
12	2,23	2,39	2,66
13	2,26	2,43	2,71
14	2,30	2,46	2,76
15	2,33	2,49	2,80
16	2,35	2,52	2,84
17	2,38	2,55	2,87
18	2,40	2,58	2,90
19	2,43	2,60	2,93
20	2,45	2,62	2,96
21	2,47	2,64	2,98
22	2,49	2,66	3,01
23	2,50	2,68	3,03
24	2,52	2,70	3,05
25	2,54	2,72	3,07
26	2,55	2,73	3,11
30	2,59	2,78	3,19

Критические значения корреляции $r_{крит}$ для уровня значимости α и степени свободы f

f \ α	0,1	0,05	0,01
1	0,988	0,997	0,999
2	0,900	0,950	0,990
3	0,805	0,878	0,959
4	0,729	0,811	0,917
5	0,669	0,754	0,874
6	0,622	0,707	0,834
7	0,582	0,666	0,798
8	0,549	0,632	0,765
9	0,521	0,602	0,735
10	0,497	0,576	0,708
11	0,476	0,553	0,684
12	0,457	0,532	0,661
13	0,441	0,514	0,641
14	0,426	0,497	0,623
15	0,412	0,482	0,606
16	0,400	0,468	0,590
17	0,389	0,455	0,575
18	0,378	0,444	0,561
19	0,369	0,433	0,549
20	0,360	0,423	0,537
25	0,323	0,381	0,487
30	0,296	0,349	0,449
35	0,275	0,325	0,418
40	0,257	0,304	0,393
45	0,243	0,287	0,372
50	0,231	0,273	0,354
60	0,211	0,250	0,325
70	0,195	0,232	0,302
80	0,183	0,217	0,283
90	0,173	0,205	0,267
100	0,164	0,196	0,254

Приложение 4

Значения коэффициента корреляции рангов Спирмена для уровня значимости α и числа измерений n

n \ α	0,1	0,05	0,01
4	0,800		
5	0,800	0,900	
6	0,771	0,829	0,943
7	0,679	0,745	0,893
8	0,619	0,714	0,857
9	0,583	0,683	0,817
10	0,552	0,636	0,782
11	0,527	0,609	0,746
12	0,496	0,580	0,727
13	0,478	0,555	0,698
14	0,459	0,534	0,675
15	0,443	0,518	0,654
16	0,426	0,500	0,632
17	0,412	0,485	0,615
18	0,399	0,472	0,598
19	0,390	0,458	0,582
20	0,379	0,445	0,568
21	0,369	0,435	0,554
22	0,360	0,424	0,543
23	0,352	0,415	0,531
24	0,344	0,406	0,520
25	0,336	0,398	0,510
26	0,330	0,389	0,500
27	0,324	0,382	0,492
28	0,318	0,375	0,483
29	0,311	0,368	0,474
30	0,306	0,362	0,466

Приложение 5

Значения *F*-критерия для уровня значимости $\alpha = 0,05$ и числа степеней свободы *f*

Знаменатель: степень свободы *f*	Числитель: степень свободы *f*
	1 2 3 4 5 6 8 12 20 30
1	161,45 199,5 215,71 224,58 230,16 234,00 238,90 243,91 248,01 250,10
2	18,51 19,00 19,16 19,25 19,30 19,33 19,37 19,41 19,45 19,46
3	10,13 9,55 9,28 9,12 9,01 8,94 8,86 8,74 8,66 8,62
4	7,71 6,94 6,59 6,39 6,26 6,16 6,04 5,91 5,80 5,75
5	6,61 5,79 5,41 5,19 5,05 4,95 4,82 4,68 4,56 4,50
6	5,99 5,14 4,76 4,53 4,39 4,28 4,15 4,00 3,87 3,81
8	5,32 4,46 4,07 3,84 3,69 3,58 3,44 3,28 3,15 3,08
10	4,96 4,10 3,71 3,48 3,33 3,22 3,07 2,91 2,77 2,70
12	4,75 3,88 3,49 3,26 3,11 3,00 2,85 2,69 2,54 2,47
20	4,35 3,49 3,10 2,87 2,71 2,60 2,45 2,28 2,12 2,04
30	4,17 3,32 2,92 2,69 2,53 2,42 2,27 2,09 1 ,93 1,84

Приложение 6

Критические значения R^2 для уровня значимости α, числа переменных (аргументов) *k* и количества опытов *n*

Уровень значимости		α = 0,1			α = 0,05			α = 0,01		
Число переменных *k*		1	2	3	1	2	3	1	2	3
Число опытов *n*	3	0,976			0,994			1,000		
	4	0,810	0,990		0,902	0,997		0,980	1,000	
	5	0,649	0,900	0,994	0,771	0,950	0,998	0,919	0,990	1,000
	6	0,532	0,785	0,932	0,658	0,864	0,966	0,841	0,954	0,993
	7	0,448	0,684	0,844	0,569	0,776	0,903	0,765	0,900	0,967
	8	0,386	0,602	0,759	0,499	0,698	0,832	0,696	0,842	0,926
	9	0,339	0,536	0,685	0,444	0,632	0,764	0,636	0,785	0,879
	10	0,302	0,482	0,622	0,399	0,575	0,704	0,585	0,732	0,830
	11	0,272	0,438	0,568	0,362	0,527	0,651	0,540	0,684	0,784
	12	0,247	0,401	0,523	0,332	0,486	0,604	0,501	0,641	0,740
	13	0,227	0,369	0,484	0,306	0,451	0,563	0,467	0,602	0,700
	14	0,209	0,342	0,450	0,283	0,420	0,527	0,437	0,567	0,663
	15	0,194	0,319	0,420	0,264	0,393	0,495	0,411	0,536	0,629
	16	0,181	0,298	0,394	0,247	0,369	0,466	0,388	0,508	0,598
	18	0,160	0,264	0,351	0,219	0,329	0,417	0,348	0,459	0,544
	20	0,143	0,237	0,316	0,197	0,297	0,378	0,315	0,418	0,498
	22	0,129	0,215	0,287	0,179	0,270	0,345	0,288	0,384	0,459
	24	0,118	0,197	0,263	0,164	0,248	0,317	0,265	0,355	0,426
	26	0,109	0,181	0,243	0,151	0,229	0,294	0,246	0,330	0,396
	28	0,101	0,168	0,2225	0,140	0,213	0,273	0,229	0,308	0,371
	30	0,094	0,157	0,210	0,130	0,199	0,256	0,214	0,289	0,349

Приложение 7

Значения критерия χ^2 для уровня значимости α и степени свободы f

f \ α	0,1	0,05	0,01
1	2,71	3,84	6,63
2	4,61	5,99	9,21
3	6,25	7,81	11,34
4	7,78	9,49	13,28
5	9,24	11,07	15,09
6	10,64	12,59	16,81
7	12,02	14,07	18,48
8	13,36	15,51	20,09
9	14,68	16,92	21,67
10	15,99	18,31	23,21
11	17,28	19,68	24,72
12	18,55	21,03	26,22
13	19,81	22,36	27,69
14	21,06	23,68	29,14
15	22,31	25,00	30,58
16	23,54	26,30	32,00
17	24,77	27,59	33,41
18	25,99	28,87	34,81
19	27,20	30,14	36,19
20	28,41	31,41	37,57
21	29,62	32,67	38,93
22	30,81	33,92	40,29
23	32,01	34,17	41,64
24	33,20	36,42	42,98
25	34,38	37,65	44,31
26	35,56	38,89	45,64
27	36,74	40,11	46,96
28	37,92	41,34	48,28
29	39,09	42,56	49,59
30	40,26	43,77	50,89
40	51,80	55,76	63,69
50	63,17	67,50	76,15
60	74,40	79,08	88,38
70	85,53	90,53	100,42
80	96,58	101,88	112,33
90	107,56	113,14	124,12
100	118,50	124,34	135,81

Учебное пособие

БАРАЗ Владислав Рувимович
ПЕГАШКИН Владимир Федорович

ИСПОЛЬЗОВАНИЕ MS EXCEL ДЛЯ АНАЛИЗА СТАТИСТИЧЕСКИХ ДАННЫХ

Редактор Н. *А. Чудина*

Подписано к печати 28.11.2014. Формат 60 × 90 $^{1}/_{16}$
Бумага офсетная Гарнитура «Таймс». Ризография
Усл.печ.л. 11,75. Уч.-изд. л. 12,8. Тираж 300 экз. Заказ № 1946

Редакционно-издательский отдел
ФГАОУ ВПО «Уральский федеральный университет
имени первого Президента России Б.Н.Ельцина»
Нижнетагильский технологический институт (филиал)
622031, Нижний Тагил, ул. Красногвардейская, 59

Отпечатано в РИО НТИ (филиа) УрФУ
181