Федеральное агентство по образованию
ГОУ ВПО «Уральский государственный технический университет – УПИ»

В.Р. БАРАЗ

Корреляционно-регрессионный анализ связи показателей коммерческой деятельности с использованием программы Excel

Рекомендовано методическим советом ГОУ ВПО УГТУ–УПИ
в качестве учебного пособия для студентов,
обучающихся по специальности 351300 – «Коммерция (торговое дело)».

Екатеринбург
2005

УДК 004.67 : 620.22 : 519.254
ББК 65.304.12 + 32.973 – 018.2

Рецензенты:
кафедра технологии металлов Уральского государственного лесотехнического университета (зав. кафедрой проф., д-р техн. наук Б.А.Потехин); доцент кафедры ОМД УГТУ-УПИ, канд. техн. наук С.И. Паршаков
Научный редактор: проф., д-р. техн. наук Б.Е. Хайкин

Бараз В.Р.
Корреляционно-регрессионный анализ связи показателей коммерческой деятельности с использованием программы Excel : учебное пособие / В.Р. БАРАЗ. – Екатеринбург : ГОУ ВПО «УГТУ–УПИ», 2005. – 102 с.

Учебное пособие предназначено для приобретения навыков применения программы Excel при выполнении цикла домашних заданий по темам «Корреляция и регрессия», «Множественная регрессия», «Непараметрические показатели связи», «Анализ хи-квадрат». Рекомендовано для студентов специальности 351300 – «Коммерция (торговое дело) в металлургии», а также для студентов других инженерных и экономических специальностей, изучающих соответствующие разделы курсов «Статистика» и «Организация эксперимента».

Библиогр. 6. Рис. 21. Табл. 14.

Подготовлено кафедрой «Металловедения»

Оглавление

Введение

Статистика необходима для того,
чтобы знать, для того, чтобы предвидеть,
для того, чтобы действовать
и для того, чтобы проверять.
(*Робер Дюма*)

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

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 другую 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 наблюдается 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 появляется 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) и ее 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 ее *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, без наложения второстепенных, то все отдельные случаи были бы совершенно одинаковы, и не возникло бы нужды анализировать всю их массу. Достаточно было бы исследовать 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 одна капля воды похожа на другую. Проводят анализ одной пробы воды и на его основе делают обобщение относительно химического состава воды. Аналогично проводятся исследования в биологии или анатомии. Например, анализируется 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 является результатом действия *основных* причин, случайный элемент – следствие действия *побочных* причин (их сочетание и действие проявляются поразному в каждом отдельном случае).

Статистическая закономерность проявляется 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 относительно точно предвидеть общий объем спроса, его структуру за год, в 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*, быть одинакового качества. Это означает, что *все элементы массива* попадают под действие *одних и тех же основных причин*. В противном случае могут возникнуть иные основные факторы, и тогда выявить общую картину 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 др.).

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

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 и аналитическую статистику. ***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 основании 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 оптимальные решения. При этом статистика отнюдь не отвергает опыт и интуицию исследователя. Ее можно рассматривать как один из компонентов процесса принятия решения, но отнюдь не как весь процесс. Поэтому есть основания считать, что статистика дополняет, но не заменяет деловой опыт, здравый смысл и интуицию человека.

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

Мощным инструментальным средством при выполнении статистических исследований является компьютерная техника. В этой связи широкое распространение в деловой сфере (точней – 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 с большим числом статистических функций.

Основное назначение данного учебного пособия – познакомить студентов с поразительными возможностями этого весьма полезного приложения и показать, как удобно его применять для выполнения достаточно стандартных статистических расчетов в деловой сфере. Таким образом, оно адресовано прежде всего студентам, обучающимся по специальности «Коммерция (торговое дело)». Вместе с тем методический способ изложения материа-

ла, приводимые практические примеры носят достаточно общий характер, и поэтому данное пособие может оказаться пригодным для студентов и других специальностей, изучающих в соответствующих учебных дисциплинах методы статистического анализа данных.

Настоящее учебное пособие можно рассматривать как определенное продолжение ранее изданного пособия по этой же теме (В.Р. Бараз. Применение программы Excel для статистических расчетов в материаловедении. – Екатеринбург : ГОУ ВПО УГТУ-УПИ, 2003. – 46 с.). Там основное внимание было уделено рассмотрению способов использования Excel для первичной статистической обработки результатов измерения, аналитического и графического описания результатов эксперимента. В данном же пособии предполагается ознакомить студентов главным образом с приемами оценки корреляционно-регрессионной зависимости, включая множественную регрессию, ранговые зависимости, поиск закономерностей для качественных данных (анализ «*хи-квадрат*»).

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

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

1. 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 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, медиана, мода), 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. Основное назначение такого подхода – возможность выявления *взаимосвязи* между параметрами.

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.

1.1. Типы зависимостей

Зависимость одной случайной величины от значений, которые принимает другая случайная величина (физическая характеристика), 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 количествен-

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

ной 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 изображение изложенных соображений представлено на рис.1.

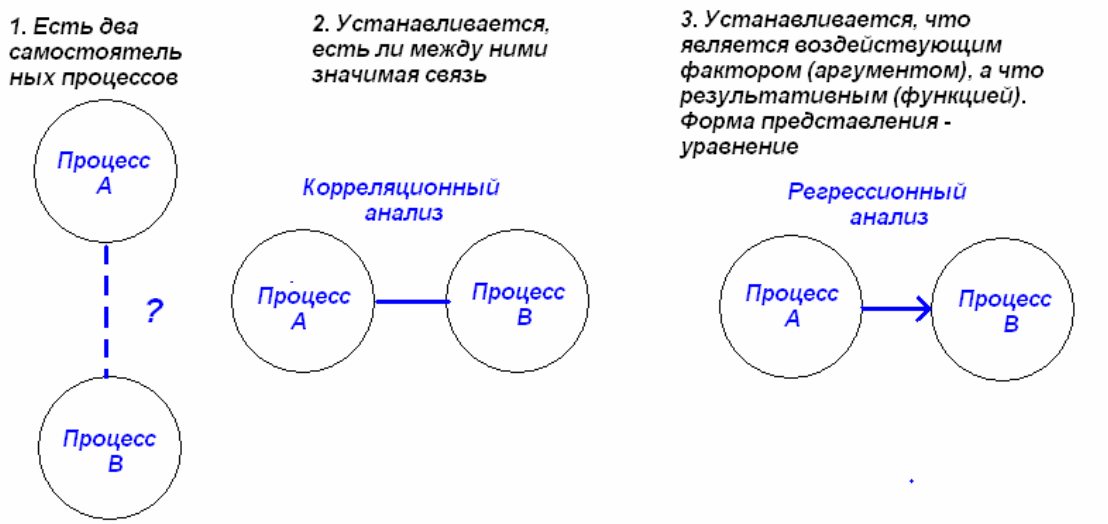


Рис.1. Схематическое пояснение сути корреляционного и регрессионного анализов

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.

Графически это (при наличии линейной зависимости) может быть представлено в виде прямой линии (рис.2*а*).

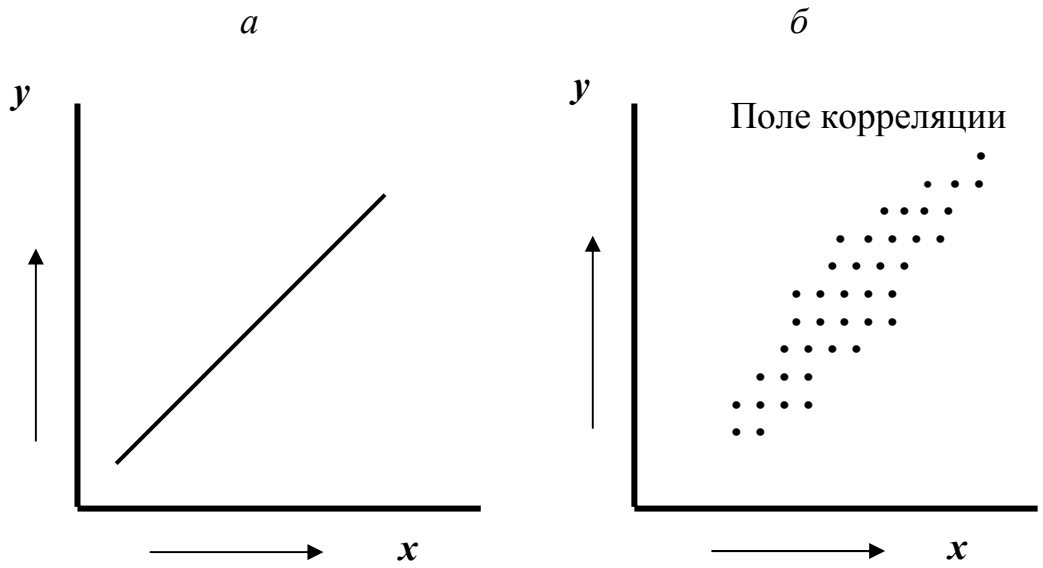


Рис.2. Зависимость функциональная (*а*) и статистическая (*б*)

Аналитически функциональная зависимость представляется в следующем виде: $y = f(x)$.

В случае статистической связи значению одного фактора соответствует какое-то приближенное значение исследуемого параметра, его точная величина является непредсказуемой, непрогнозируемой, 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 объема товарооборота. В этой связи помимо факторного признака x (объема 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*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 линейная взаимосвязь, которая 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.

1.2. Методы определения корреляционной связи

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

Если между парами совокупностей просматривается вполне очевидная связь (ранее нами это исследовалось, 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 наличие связи *качественно*, 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 и 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.

1.3. Расчет коэффициента парной корреляции и его статистическая проверка

Существуют различные аналитические приемы определения коэффициента *r*. Известна такая формула:

$$r = \frac{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{S_x S_y},$$

где S_x и S_y – среднеквадратичное отклонение соответственно для каждого рассматриваемого массива чисел; x_i и y_i – текущие значения единиц обеих совокупностей; $\bar{x}$ и $\bar{y}$ – их средние величины и n – число измерений (элементов) в каждой совокупности.

В литературе по статистике рекомендуется использовать также и другое выражение:

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

В этом случае отпадает необходимость вычислять отклонения текущих (индивидуальных) значений от средней величины. Это исключает ошибку в расчетах при округлении средних величин.

Зная коэффициент корреляции, можно дать качественно-количественную оценку тесноты связи. Используются, например, специальные табличные соотношения (так называемая шкала Чеддока).

Ее представление может иметь следующий вид (табл. 1):

Таблица 1

Качественная оценка тесноты связи

Величина коэффициента парной корреляции	Характеристика силы связи
До 0,3	Практически отсутствует
0,3–0,5	Слабая
0,5–0,7	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 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 (*КПК*) 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 $r_{\text{крит}}$, значения которого отыскиваются из специальной таблицы.

Если окажется, что $|r_{\text{расч}} \geq r_{\text{крит}}|$, 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 $|r_{\text{расч}} < r_{\text{крит}}|$, делается заключение об отсутствии значимой связи.

Перейдем 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, 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.

Таблица 2

Снижение эффективности мышиной охоты кота Матроскина и ее возможные причины

Дни	Число пойманных мышей	Средняя дневная температура, °С	Продолжительность 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:

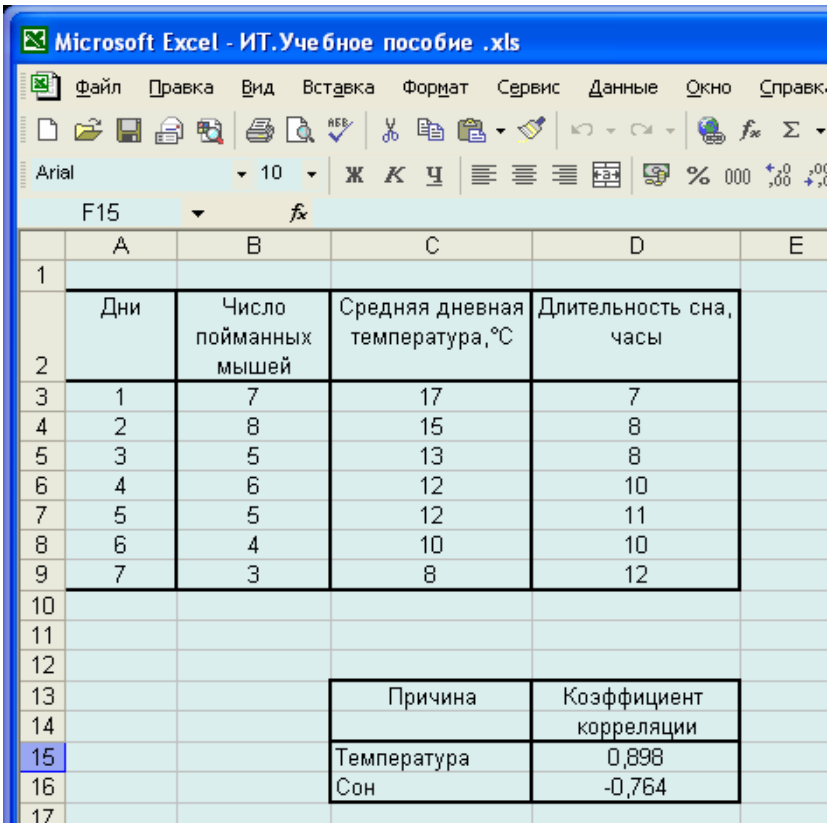
– нажмем кнопку **Пуск** в панели задач (находится слева на самой нижней полосе **Рабочего стола**), а затем откроем во всплывающем меню опцию **Программы**;

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

Подготовим табл.1 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 «Продолжительность сна, часы». Затем разместим сами числовые наборы соответственно в диапазонах ячеек A3:A9, B3:B9, C3:C9 и D3:D9 (рис.2).

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 (рис.3).

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					
15			Температура	0,898	
16			Сон	-0,764	
17					

Рис.3. Исходные данные и расчет коэффициента корреляции.

Действуем в такой последовательности:

– в итоговой таблице активизируем ячейку D15, куда и будет помещено первое расчетное значение *КПК*;

– запустим **Мастер функций** (ищем в инструментальной строке значок ***f***) и в всплывающем диалоговом окне укажем требуемую категорию – **Статистические**, а затем выделим нужную функцию **Коррел**, после чего – **ОК** (рис.4);

– 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.5);

– и, наконец, нажмем кнопку **ОК**.

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

В выделенных ячейках D15 и D16 (рис.3) появятся числа, указывающие соответствующие значения коэффициентов корреляции. После установления нужной разрядности в окончательном виде получим следующие значения: $r_{расч1} = 0{,}898$ и $r_{расч2} = -0{,}764$.

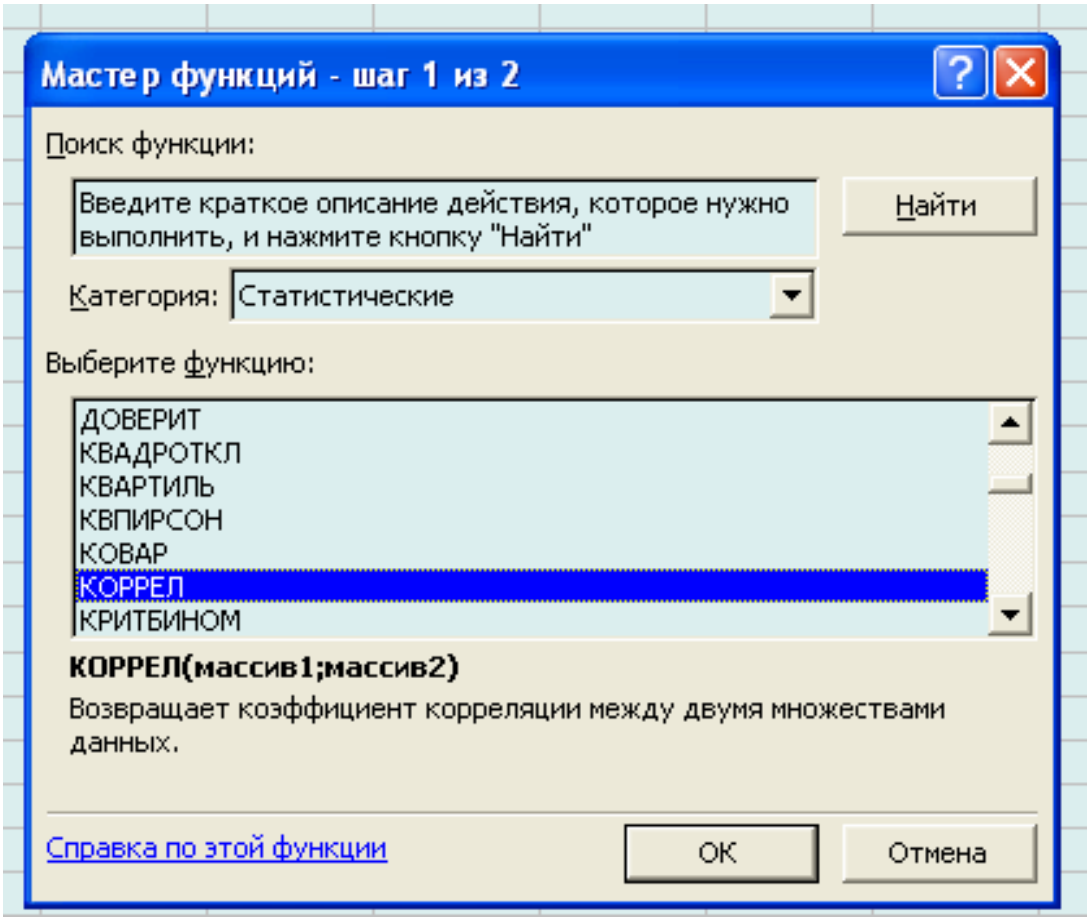


Рис.4. Диалоговое окно **Мастер функций**

Первый коэффициент показывает, насколько заметна теснота связи параметров «Количество пойманных мышей – средняя дневная температу-

ра». Второй показатель характеризует другую изучаемую связь «Количество пойманных мышей – продолжительность 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).

Теперь надлежит дать статистическую оценку выполненных нами расчетов, т.е. проверить на адекватность рассматриваемые события. Для этого сопоставим расчетные значения коэффициентов $r_{\text{расч}}$ с табличным показателем $r_{\text{крит}}$. Используя *прил. 1*, находим, что для уровня значимости (т.е. вероятности допустимой ошибки в прогнозе) $\alpha = 0{,}05$ и заданного числа измерений n табличное значение $r_{\text{крит}} = 0{,}754$.

Как видно, в обоих случаях выполняется соотношение $|r_{\text{расч}} \geq r_{\text{крит}}|$, 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.

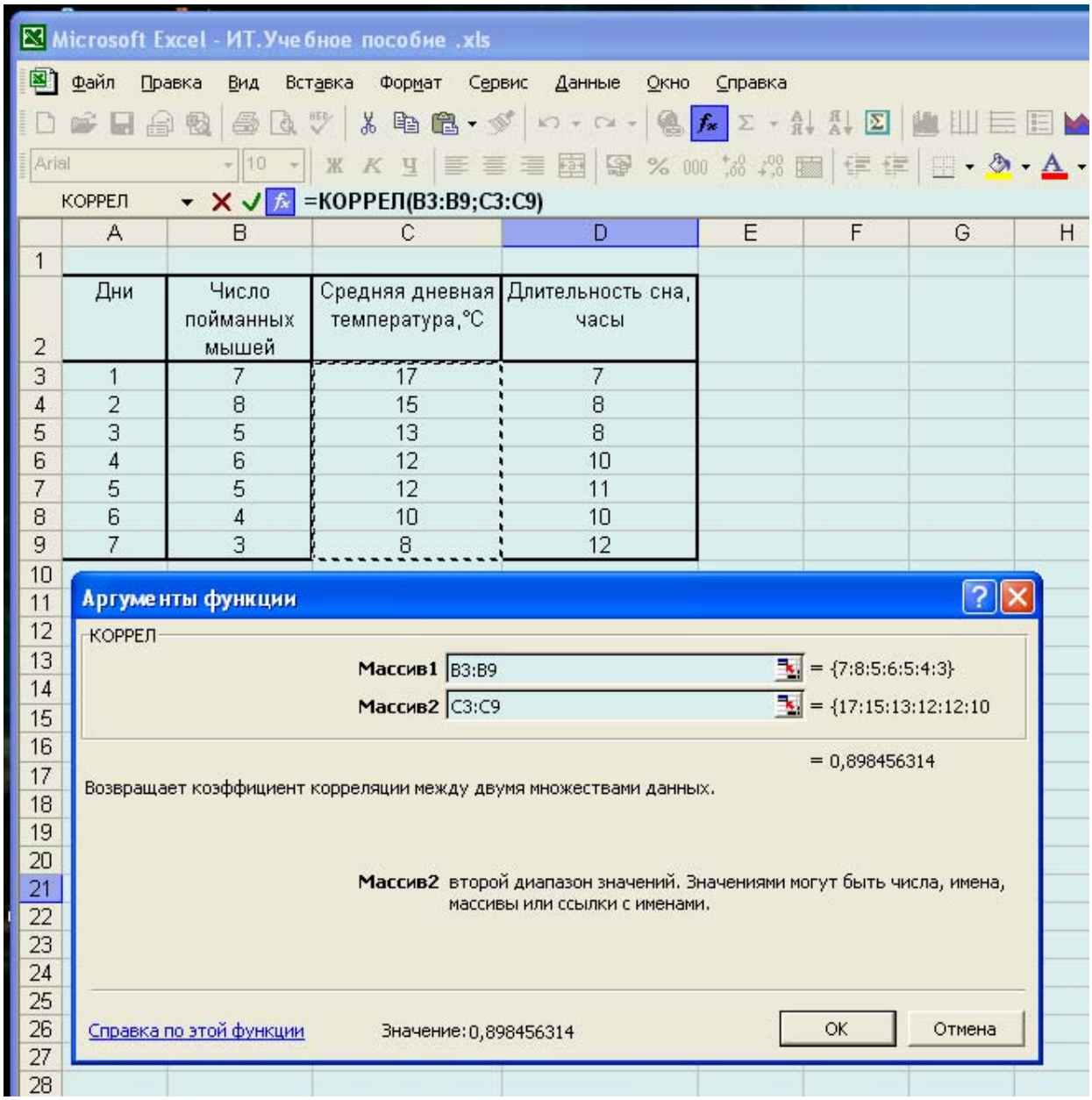


Рис.5. Диалоговое окно ввода параметров корреляции.

Примечание. Заметим, что в таблице для $r_{крит}$ (*прил.1*) вместо привычных значений числа измерений n стоит показатель f, характеризующий так называемую *степень свободы*. Число степеней свободы в статистике определяется как разность 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. $n - 2 = 7 - 2 = 5$.

1.4. 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 прибыль).

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. И не более того.

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 – высокой потребности в продукции фирмы, что и послужило причиной и приема на работу новых менеджеров, и создания новых производственных мощностей.

В истории статистики известен один классический 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 коэффициент парной корреляции. 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*) R^2, который равен квадрату коэффициента корреляции (r^2). Он показывает, в какой мере изменчивость y (результативного признака) объясняется поведением x (факторного признака), или иначе: какая часть общей изменчивости y вызвана собственно влиянием x. Этот показатель вычисляется путём простого возведения в квадрат коэффициента корреляции. Тем самым доля изменчивости y, определяемая выражением $1 - R^2$, оказывается необъясненной.

Допустим к примеру, что коэффициент корреляции совокупности данных, относящихся к производственным затратам, равняется 0,869193. Следовательно, значение R^2 равно

$$R^2 = 0{,}869193^2 = 0{,}755 \text{ или } 75{,}5\ \%.$$

Это значение R^2 говорит о том, что 75,5 % вариации (изменчивости), скажем, недельных затрат объясняется количеством изделий, выпущенных за

неделю. Остальная часть (24,5 %) 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.

1.5. 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 уве-

личению, 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 d_i.

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

Выборки (их объем): 1, 2, 3, …, n

1-я совокупность (признак x): $x_1,\ x_2\ , x_3,\ \ldots, x_n$

2-я совокупность (признак y): $y_1,\ y_2,\ y_3,\ \ldots, y_n$.

Здесь оба параметра x и y принимают только целочисленные значения в количестве, равном n.

Тогда формула коэффициента корреляции рангов Спирмена (этот коэффициент именуют p) имеет следующий вид:

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

Рассмотрим определение этого коэффициента на следующем примере.

Студенты третьего курса, обучающиеся по специальности «Коммерция (торговое дело)», проходили производственную практику в качестве стажеров на двух фирмах, занимающихся торгово-закупочными операциями с цветными металлами. Число студентов составляло 12 человек. Они работали вначале в течение двух недель на фирме «Колокольный звон», занимающейся в основном изделиями из бронзы, а остальные две недели – на фирме «Мельхиор», коммерческий интерес которой преимущественно был направлен на торговлю декоративно-ювелирными изделиями из медноникелевых сплавов. Получив жалование, заработанное усердным трудом, ребята решили выяснить, отличаются ли принципиально их материальные успехи в

зависимости от того, где они приобретали практические навыки своей будущей профессии.

Эту задачу мы постараемся решить двумя приемами. Вначале выполним необходимые расчеты «вручную», проведя все необходимые рутинные операции с использованием вспомогательных табличных материалов, а также последующих скучных расчетов. Затем решим ту же задачу, воспользовавшись помощью замечательного Excel.

Итак, в табл.3 укажем условные порядковые номера студентов, их заработок (*тыс. руб.*) на каждой фирме, соответствующее условное место (ранг), который они занимают в зависимости от размера заработка, а также все необходимые вспомогательные выкладки.

Как видно из результатов сопоставления рангов материальных достижений студентов, их фактические показатели выглядят достаточно пестро. В одних случаях ранги были вполне совпадающими (например, у студентов под номерами 2, 8 и 12, но особенно полное совпадение у студентов с номерами 7 и 10), в других же заметно различались (например, у
студентов под номерами 3, 5, 6 и 11). Возникает вопрос: насколько точно можно было прогнозировать успешную (или, напротив, менее удачную) работу студентов в указанных фирмах? Для ответа вычислим коэффициент корреляции рангов Спирмена, используя результаты расчетов в графе 7:

$$p = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \times 105}{12(12^2-1)} = 0{,}633 .$$

Как и линейный коэффициент корреляции, коэффициент корреляции рангов может также меняться от –1 до +1. Используя шкалу Чеддока, по результатам расчетов коэффициента Спирмена можно предположить наличие заметной прямой зависимости между итогами работы студентов на данных фирмах. Однако следует учесть, что ранговый показатель был рассчитан по небольшому объему исходной информации (n = 12). Не является ли отличие

рангового коэффициента от нуля лишь следствием случайных совпадений результатов деятельности студентов на обеих фирмах? Иначе говоря, указанные совпадения не есть результат влияния каких-то иных факторов (дурной характер работодателя, финансовое положение фирмы, знойная жара в этот период лета и проч.), а всецело определяются усердием самих студентов.

Чтобы ответить на этот вопрос более определенно, оценим статистическую значимость расчетного коэффициента. Для этого его значение $p_{\text{расч}}$ нужно сопоставить с критическими (табличными) $p_{\text{табл}}$. Используется таблица, напоминающая таблицу *t*-критерия (*прил.2*).

Найдем табличное значение коэффициента $p_{\text{табл}}$, для $\alpha = 0{,}05$ и $n = 12$ его величина составит 0,580. Поскольку $p_{\text{расч}} > p_{\text{табл}}$ (0,633 и 0,580), то с вероятностью 95 % можно утверждать, что исследуемая связь является значимой. Однако для уровня значимости $\alpha = 0{,}01$ табличное значение $p_{\text{табл}} = 0{,}723$. Тем самым уже для вероятности 99 % наличие связи становится неочевидной.

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

Таблица 3

Расчетная таблица для определения коэффициента корреляции рангов Спирмена

Порядковый номер студента	Заработок, фирма «Колокольный звон», *тыс. руб.* (x)	Заработок, фирма «Мельхиор», *тыс. руб.* (y)	Ранг R_x	Ранг R_y	Разность рангов $d = \lvert R_x - R_y \rvert$	d^2
1	*2*	*3*	*4*	*5*	*6*	*7*
1	2,8	3,3	6	3	3	9
2	3,1	3,0	5	6	1	1
3	2,0	2,8	11	7	4	16
4	3,2	4,1	4	2	2	4
5	2,4	2,1	8	12	4	16
6	3,3	2,7	3	8	5	25
7	2,2	2,5	10	10	0	0
8	1,8	2,3	12	11	1	1
9	2,5	3.2	7	4	3	9
10	2,3	2.6	9	9	0	0
11	3,5	3,1	1	5	4	16
12	3,4	4,5	2	1	1	1
Итого						105

Заметим, что ранговый коэффициент корреляции Спирмена 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 важно).

Теперь ту же задачу мы решим, используя компьютерные расчеты. В данном случае Excel нам поможет выполнить рутинные расчеты, хотя сама процедура поиска коэффициента корреляции Спирмена будет носить схожий характер.

Итак, запустим программу Excel. В открывшемся рабочем листе Excel (**Лист 1**) сформируем исходную таблицу, в которой поместим данные, соответствующие содержимому колонок 1-3 табл.3. Эта таблица будет располагаться в ячейках A1:C13. Итоговый результат представлен на рис.6.

	A	B	C
1	Порядковый номер студента	Заработок, фирма "Колокольный звон", тыс. руб.	Заработок, фирма "Мельхиор", тыс. руб.
2	1	2,8	3,3
3	2	3,1	3
4	3	2	2,8
5	4	3,2	4,1
6	5	2,4	2,1
7	6	3,3	2,7
8	7	2,2	2,5
9	8	1,8	2,3
10	9	2,5	3,2
11	10	2,3	2,6
12	11	3,5	3,1
13	12	3,4	4,5

Рис.6. Исходные данные в таблице Excel

Далее будем двигаться следующим образом: запустим опции **Сервис/Анализ данных/Ранг и персентиль.**

Примечание. Тут следует дать предварительное пояснение. В отношении рангов рассуждения у нас уже были. Теперь дадим разъяснение по поводу термина персентиль (или, как принято писать, перцентиль).

Как уже говорилось, для характеристики формы распределения вариационного ряда применяют *ранговые* показатели. Под этим понимают такие единицы исследуемого массива, которые занимают определенное место в вариационном ряду (например, десятое, двадцатое и т.д.). Они получили название *квантилей* или *градиентов*. Квантили в свою очередь подразделя-

ются на *квартили, децили* и *перцентили*. Различие между ними в том, *на какое количество частей делится вариационный ряд*. Если на 4 части – это квартили; на 10 – децили и, наконец, на 100 – перцентили.

Поясним это на примере перцентилей. ***Перцентили*** – это характеристики набора данных, которые выражают ранги элементов массива в виде процентов от 0 до 100 %, а не в виде чисел от 1 до *n*. В результате наименьшему значению соответствует нулевой перцентиль, наибольшему – 100-й перцентиль, медиане – 50-й перцентиль и т.д. Следовательно, перцентили можно рассматривать как показатели, разбивающие анализируемый массив на определенные части.

Заметим, что перцентиль представляет собой какой-то элемент массива, имеющий определенный ранг и выраженный в тех же единицах, что и сам массив в целом. Так, 60-й перцентиль эффективности сбора металлолома в конторе «Ржавая подкова» составляет, скажем, 85062 руб. (измерен не в процентах, а в рублях, как элемент набора данных). Если этот 60-й перцентиль, равный 85062 руб., характеризует деятельность определенного агента по заготовкам (например, г-на Пупкина), то это означает, что примерно 60 % других тружеников имеют результат ниже, чем у г-на Пупкина, а 40 % – более высокие показатели.

Перцентили используются для двух целей:

– чтобы показать значение элемента в массиве при заданном перцентильном ранге (например, «10-й перцентиль равен 46293 руб.»);

– чтобы показать перцентильный ранг значения данного элемента в рассматриваемом массиве статистических данных (например, «эффективность заготовок металлолома агента г-на Козлевича составляет 65994 руб., что соответствует 55-му перцентилю»).

Продолжим рассмотрение нашей задачи. В диалоговом окне **Ранг и персентиль** заполним поле **Входной интервал** (рис.7).

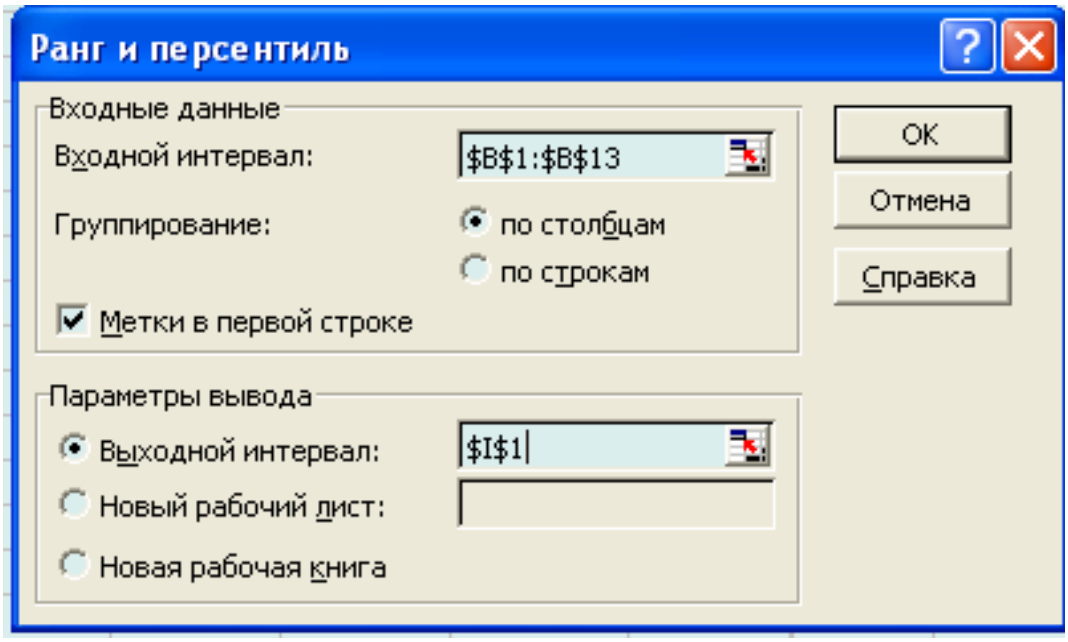


Рис. 7. Диалоговое окно ***Ранг и перцентиль***

В нем укажем данные 2-й графы табл. 3 (вместе с заголовком), относящиеся к фирме «Колокольный звон» (это диапазон ячеек $B$1:$B$13). Отметим флажком позицию **Метки в первой строке** (поскольку нам нужно сохранить заголовок этой графы), а затем в окне **Выходной интервал** укажем ячейку $I$1, в которой будет размещена таблица с рассчитанными показателями рангов и перцентилей. После этого – кнопка **ОК**.

Затем аналогичным образом поступим с данными 3-й графы (сведения от фирмы «Мельхиор»). При заполнении диалогового окна **Ранг и персентиль** отметим диапазон ячеек $C$1:$C$13, а для опции **Выходной интервал** покажем ячейку, которая должна быть по соседству с первой половинкой нашей общей таблицы. Это ячейка $M$1.

В окончательном виде наша таблица примет следующий вид (рис.8).

Как видно, Excel аккуратно проранжировал результаты по каждому эпизоду, расположив студентов по местам в соответствии с их материальны-

ми успехами, а также указал их перцентильный ранг (в %). Для дальнейших рассуждений данные по перцентилям мы использовать не станем, а вот ранги окажутся совершенно необходимыми.

I	J	K	L	M	N	O	P
Точка	*Фирма "Колокольный звон"*	*Ранг*	*Процент*	*Точка*	*Фирма "Мельхиор"*	*Ранг*	*Процент*
11	3,5	1	100,00%	12	4,5	1	100,00%
12	3,4	2	90,90%	4	4,1	2	90,90%
6	3,3	3	81,80%	1	3,3	3	81,80%
4	3,2	4	72,70%	9	3,2	4	72,70%
2	3,1	5	63,60%	11	3,1	5	63,60%
1	2,8	6	54,50%	2	3	6	54,50%
9	2,5	7	45,40%	3	2,8	7	45,40%
5	2,4	8	36,30%	6	2,7	8	36,30%
10	2,3	9	27,20%	10	2,6	9	27,20%
7	2,2	10	18,10%	7	2,5	10	18,10%
3	2	11	9,00%	8	2,3	11	9,00%
8	1,8	12	,00%	5	2,1	12	,00%

Рис. 8. Расчетная таблица с показателями рангов и перцентилей

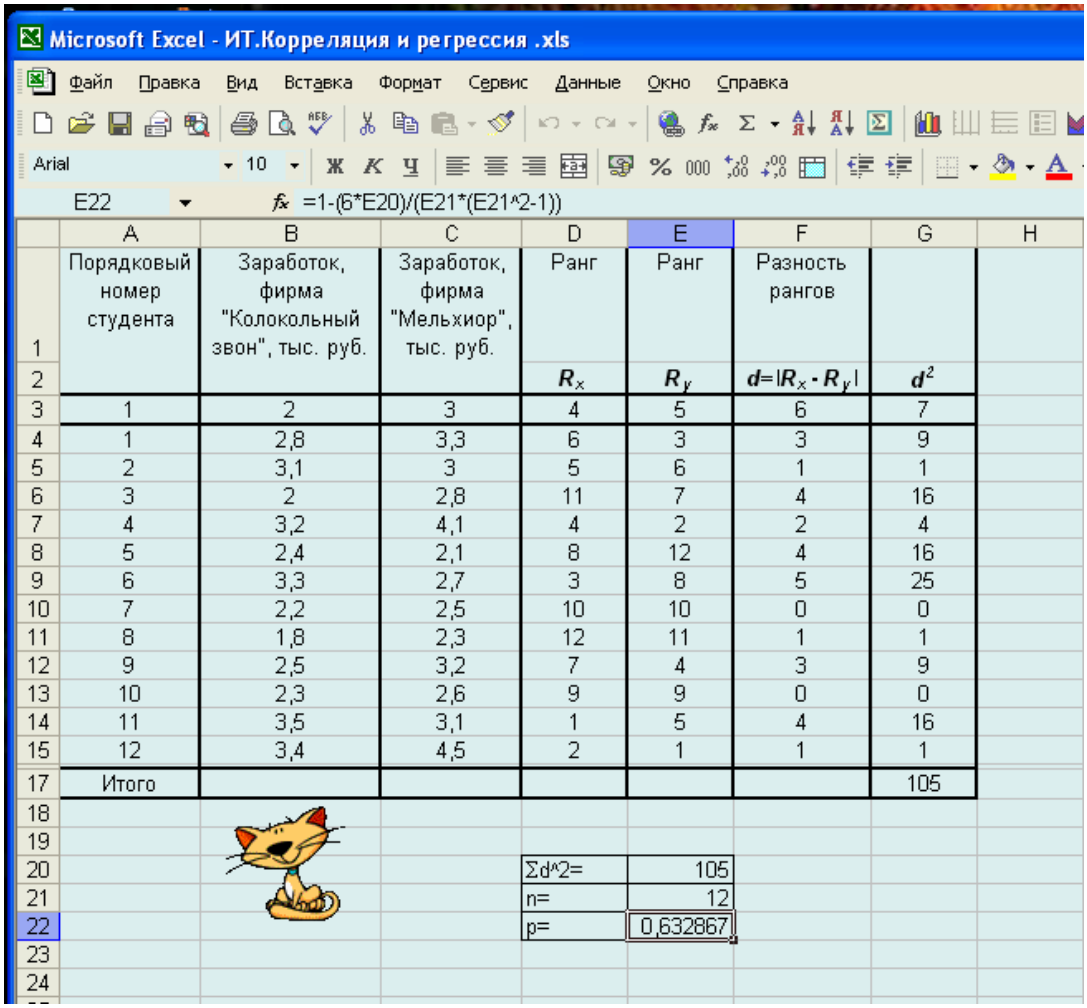
На основании ранговых оценок организуем сводную таблицу, аналогичную уже знакомой нам табл.3 (рис.9). Для удобства перейдем на другой рабочий лист (**Лист 2**). Для выполнения последующих расчетов используем итоговый результат, отражающий сумму разностей квадратов рангов, равную 105. Оформим вспомогательную таблицу (рис.9), в которой укажем значение $\Sigma d^2 = 105$, размер выборки $n = 12$, а также предусмотрим в ней ячейку, где поместим рассчитанное значение коэффициента ранговой корреляции p (ячейка E22).

Поместим курсор в ячейку E22, а затем в поле формулы запишем уравнение, по которому будем рассчитать коэффициент p. Выглядит оно так:

= 1 – 6*(E20)/(E21*(E21^2 – 1))

В ячейке появится искомый результат 0,632867. С округлением принимаем его равным 0,633 – коэффициент оказался именно таким, каким мы его вычислили «вручную».

Полученный результат показывает, что в данной ситуации надлежит высказать совершенно те же соображения по поводу исследуемого процесса, какие были сделаны для случая расчета коэффициента *p* традиционным способом. При доверительной вероятности 0,95 студенты вполне могут горделиво полагать, что их материальные достижения всецело определяются личным усердием и не зависят от каких-то иных привходящих факторов. Однако требование более строгой оценки (с вероятностью 99 %) делает такое мнение менее очевидным и для значимого статистического вывода возникает необходимость расширить выборку (привлечь для анализа большее число студентов) либо (при невозможности это сделать) отнестись к результату вполне философски.



	A	B	C	D	E	F	G	H
1	Порядковый номер студента	Заработок, фирма "Колокольный звон", тыс. руб.	Заработок, фирма "Мельхиор", тыс. руб.	Ранг	Ранг	Разность рангов		
2				R_x	R_y	$d=\|R_x - R_y\|$	d^2	
3	1	2	3	4	5	6	7	
4	1	2,8	3,3	6	3	3	9	
5	2	3,1	3	5	6	1	1	
6	3	2	2,8	11	7	4	16	
7	4	3,2	4,1	4	2	2	4	
8	5	2,4	2,1	8	12	4	16	
9	6	3,3	2,7	3	8	5	25	
10	7	2,2	2,5	10	10	0	0	
11	8	1,8	2,3	12	11	1	1	
12	9	2,5	3,2	7	4	3	9	
13	10	2,3	2,6	9	9	0	0	
14	11	3,5	3,1	1	5	4	16	
15	12	3,4	4,5	2	1	1	1	
17	Итого						105	
18								
19								
20				Σd^2=	105			
21				n=	12			
22				p=	0,632867			
23								
24								

Рис. 9. Фрагмент рабочего листа Excel с обобщенной таблицей и данными для расчета коэффициента корреляции Спирмена.

Регрессионный метод оценки коммерческой деятельности

Если мой сосед бьет жену каждый день, а я никогда,
то с точки зрения статистики мы оба бьем своих жен через день.

(Бернард Шоу)

Подожди - и плохое само собой исчезнет...,
нанеся положенный ущерб.

(Расширенный закон Мэрфи)

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 – одна из них является *аргументом*, а вторая *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 явля-

ется 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.

2.1. 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.10).

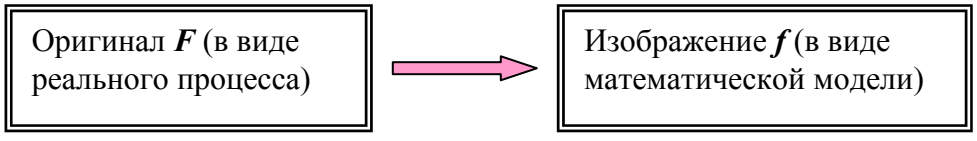


Рис.10. Схематическая связь между оригиналом и моделью объекта

Задача аппроксимации часто возникает при обработке результатов экспериментов, когда становится необходимым подобрать математическую модель изучаемого процесса, т.е. дать его аналитическое описание в виде так называемой *эмпирической формулы*.

При подборе эмпирической формулы 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.

2.2. 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 ус-

пешно решается современными средствами информационных технологий (программа Excel).

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

Термином ***линейный регрессионный анализ*** обозначают такое прогнозирование, которое описывается линейной взаимосвязью между исследуемыми переменными: $y = b_0 + b_1 x$.

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

гиперболическая	$y = b_0 + b_1 / x$;
показательная	$y = b_0 + b_1^{x}$;
степенная	$y = b_0 x^{b1}$;
параболическая	$y = b_0 + b_1 x + b_2 x^2$;
логарифмическая	$y = b_0 + b_1 \lg x$;
экспоненциальная	$y = b_0 exp\,(b_1 x)$ и другие.

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

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:

$$y = f(x) = b_0 + b_1 x + b_2 x^2 + \ldots + b_m x^m ,$$

где b_0 , b_1 , b_2,…, 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 опытным данным.* 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.

2.3. Метод наименьших квадратов

Для определения коэффициентов уравнения регрессии b применяют разные методы (графический, метод средних), однако наибольшее распространение получил метод наименьших квадратов (*МНК*).

Пусть обсуждается некоторая зависимость $y = f(x)$, 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 такого разброса 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 (теоретическими показателями) должна быть наименьшей.

Довольно часто при описании аппроксимирующей функции ограничиваются простым видом полиноминальной зависимости, полагая ее линейной, т.е. в виде уравнения прямой $y = b_0 + b_1 x$. Здесь *свободный член* b_0 характеризует *сдвиг* и равен тому значению y, которое получается при $x = 0$, а *коэффициент* b_1 определяет *наклон* линии.

Отыскание коэффициентов b_0 и b_1 осуществляется по *МНК*.

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

Предположим, что экспериментальные точки на графике укладываются так, что по ним вполне возможно провести прямую линию (рис.11). Значения функции $\hat{y}_i$ в этом случае можно записать в виде линейного уравнения: $\hat{y}_i = b_0 + b_1 x_i$. Расстояние по ординате (вертикали) от точки y_i до прямой составит: $b_0 + b_1 x_i - y_i = \varepsilon_i$, где $b_0 + b_1 x_i = \hat{y}_i$ – *рассчитанное (теоретическое)* значение функции; y_i – ее *измеренное (опытное)* значение и ε_i – *разница (расстояние)* между $\hat{y}_i$ и y_i.

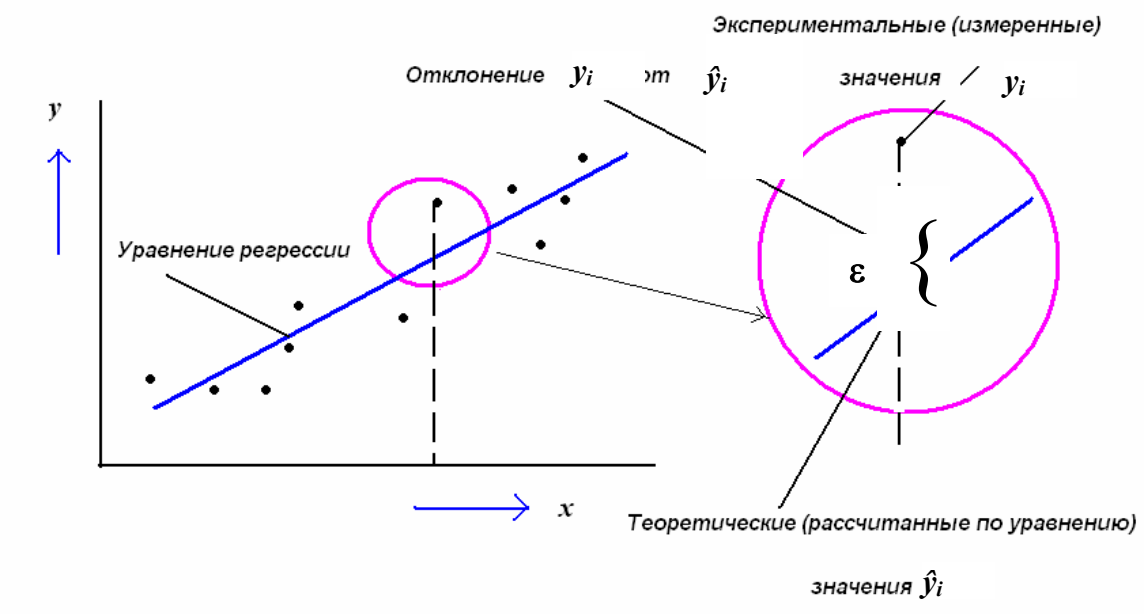


Рис. 11.Схематическое пояснение содержания метода наименьших квадратов

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

Минимум этой суммы ищется по правилам дифференциального исчисления. В результате для определения b_0 и 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, экспонент) 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; … x_n экспериментальные результаты для y. Подобную оценку принято называть проверкой на значимость или адекватность.

2.4. 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 ее сплавов. При этом г-на Тютякина интересовало, сколько медного металлолома в среднем можно собрать за одну рабочую смену (8 часов).

Боб Деканкин, знакомый с методом регрессионного анализа, решил взяться за порученное дело. В течение месяца он аккуратно регистрировал результаты сбора медного металлолома. Это позволило ему представить в табличной форме (табл.4) основные итоги, указав для статистического массива n = 8: а) затраченное время (часы) и б) вес собранного металлолома (кг).

Таблица 4

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.12).

2.4.1. Использование традиционных способов расчета

На первом этапе проведем вычисление традиционным, а потому и самым утомительным способом, т.е. «вручную». Здесь нам в лучшем случае может помочь лишь калькулятор.

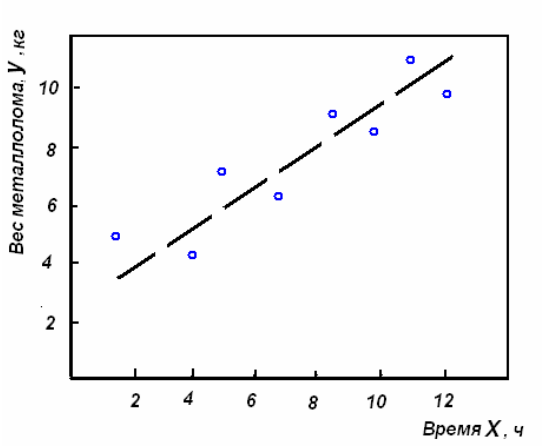


Рис.12. Графическое изображение исследуемой зависимости $y = f(x)$

Вычисление коэффициентов регрессии удобнее проводить в табличной форме. Для этого заполним табл.5, 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: $\bar{x} = 59,5/8 = 7,44$ ч и $y = 61,5/8 = 7,69$ кг.

Значения полученных сумм подставляем в формулу для последующей проверки. Получим:

$2072,00 = 541,75 + 2\times510,25 + 509,75;$

$2072,00 = 2072,00.$

Следовательно, вычисления выполнены правильно.

Таблица 5

Вспомогательная таблица для расчета коэффициентов регрессии

№ п/п	x	$У$	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:

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

$$b_0 = \frac{541,75\times61,50 - 59,50\times510,25}{8\times541,75 - 59,50^2} = 3,73 \text{ кг.}$$

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

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

Следовательно, 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*):

$$F = \frac{S^2_{\text{ад}}}{S^2_{\text{общ}}}.$$

Где $S^2_{\text{ад}}$ – *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 предсказании 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 $F_{\text{крит}}$.

Если $F_{\text{расч}} < F_{\text{крит}}$, то модель признается *адекватной*, т.е. с заданной степенью достоверности (надежности) она верно предсказывает реальный результат. Если же $F_{\text{расч}} > F_{\text{крит}}$, 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 хуже по сравнению с опытом предсказывает результат модель.*

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

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

Здесь число степеней свободы $f = n - (k + 1)$, где n – число опытов в эксперименте (т.е. объем случайной выборки); k – число изучаемых факторов.
Для однофакторного эксперимента имеем $f = n - 2$ и тогда

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

Вторая характеристика в формуле для расчета *F*-критерия (знаменатель) – это так называемая *усредненная,* или *общая дисперсия*. В качестве таковой принимается квадрат стандартной ошибки $S^2_{\text{общ}}$. 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 так:

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

Вернемся к нашему примеру. Оценим статистическую пригодность полученного линейного уравнения. Показатель $S^2_{\text{ад}}$ удобно вычислять в табличной форме (табл.6). Расчет проведем по формулам:

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

Таблица 6

Вспомогательная таблица для проверки уравнения на адекватность

№ п/п	x_i	y_i	$\hat{y}_i$=3,73+0,53x	$y_i - \hat{y}_i$	$(y_i - \hat{y}_i)^2$	$\bar{y}_i - y_i$	$(y_i - y_i)^2$
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=36,30

Определим величину критерия Фишера:

$$F_{\text{расч}} = \frac{S^2_{\text{ад}}}{S^2_{\text{общ}}} = \frac{1{,}11}{4{,}63} = 0{,}24 .$$

Определим табличное значение для $\alpha = 0{,}05$, а также степеней свободы для числителя f_1 ($S^2_{\text{ад}}$) и знаменателя f_2 ($S^2_{\text{общ}}$). Они составят соответственно $f_1 = n - 2$, т.к. $f = n - (k + 1)$, где n – число опытов в эксперименте (т.е. составляет объем случайной выборки); k – число изучаемых факторов. Для однофакторного эксперимента имеем $f = n - 2$.

Для второго показателя $f_2 = n - m$, где m – количество вычисленных констант для переменной y, которая соответствует среднемарифметическому $\bar{y}$ (т.е. $m = 1$). Тогда $f_2 = n - 1$, а $F_{\text{крит}\,(0{,}05;\, f1;\, f2)} = 3{,}87$ (*прил.3*).

Поскольку $0{,}24 < 3{,}87$, то с вероятностью 95 % можно утверждать, что рассматриваемое уравнение адекватно и 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: за 8 часов работы это составит почти 8 кг ($3{,}7 + 0{,}53 \times 8 = 7{,}97$).

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

ем для анализа используется и, следовательно, каким образом дается надлежащее заключение.

2.4.2. Расчет с использованием компьютерной программы

А теперь покажем, как всю эту громоздкую и довольно затратную по времени процедуру можно весьма элегантным образом заменить услугами Excel.Для этого на рабочем листе Excel предварительно создадим таблицу с исходными данными, в которой укажем содержимое табл.4. Причем саму таблицу построим по столбцам и поместим ее в ячейках A1:C9. Итоговый результат показан на рис.13.

Далее будем действовать привычным образом:

– в главном меню запустим серию команд **Сервис/Анализ данных/Регрессия**;

– 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);

– отметим **Уровень надежности** (доверительную вероятность), равный 95 %;

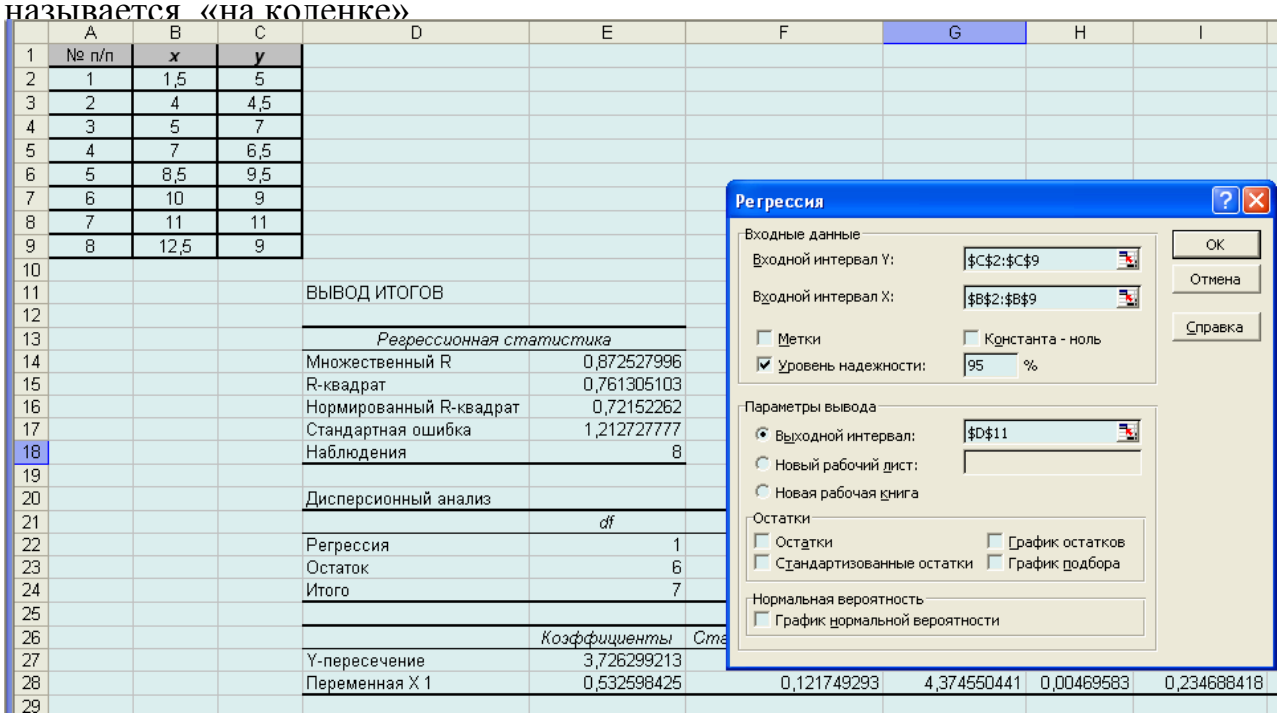
– укажем в окне вывода **Выходной интервал** ту ячейку, от которой будет формироваться весь блок получаемых статистических показателей, это D11;

– после чего нажмем кнопку **ОК**.

На рис.13 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».

Старательный Excel выдал, как мы видим, весьма богатый набор разнообразных статистических материалов. Выберем, однако, из них 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 b_0 и коэффициент b_1 при аргументе x. Если затем провести надлежащее округление до второго знака после запятой, то получим уже знакомые нам числа 3,73 и 0,53, которые были рассчитаны ранее, что называется «на коленке».



	A	B	C	D	E	F	G	H	I
1	№ п/п	*x*	*y*						
2	1	1,5	5						
3	2	4	4,5						
4	3	5	7						
5	4	7	6,5						
6	5	8,5	9,5						
7	6	10	9						
8	7	11	11						
9	8	12,5	9						
10									
11				ВЫВОД ИТОГОВ					
12									
13				*Регрессионная статистика*					
14				Множественный R	0,872527996				
15				R-квадрат	0,761305103				
16				Нормированный R-квадрат	0,72152262				
17				Стандартная ошибка	1,212727777				
18				Наблюдения	8				
19									
20				Дисперсионный анализ					
21					*df*				
22				Регрессия	1				
23				Остаток	6				
24				Итого	7				
25									
26					*Коэффициенты*	*Сте*			
27				Y-пересечение	3,726299213				
28				Переменная X 1	0,532598425	0,121749293	4,374550441	0,00469583	0,234688418
29									

Рис. 13. Лист Excel с результатами расчета коэффициентов регрессии

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 проце-

дура была выполнена нами, 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 (т.к. оценку пригодности уравнения мы дали, хотя и весьма обременительным способом), более обстоятельно с возможностями Excel познакомимся в следующей главе.

3. 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_k 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) используется также такое понятие, как *ошибка прогнозирования* Δy**.** Последняя понимается как 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_k на изменение функции *y* над ее 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_k *остаются неизменными*. Построение доверительных интервалов и проверка 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 объем ежедневной продажи молока ***y*** *(литров в день) зависит от а) присутствия среди покупателей бабушек с внучками (их доля от общего числа покупателей* $\boldsymbol{x_1}$*, %) и б) участия в коммерции пса Шарика (относительное время* $\boldsymbol{x_2}$*, когда он помогал работать за прилавком, %). Тщательные наблюдения Матроскин вел в течение 20 рабочих дней, результаты которых представил в табличной форме (табл.7). 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.

Таблица 7

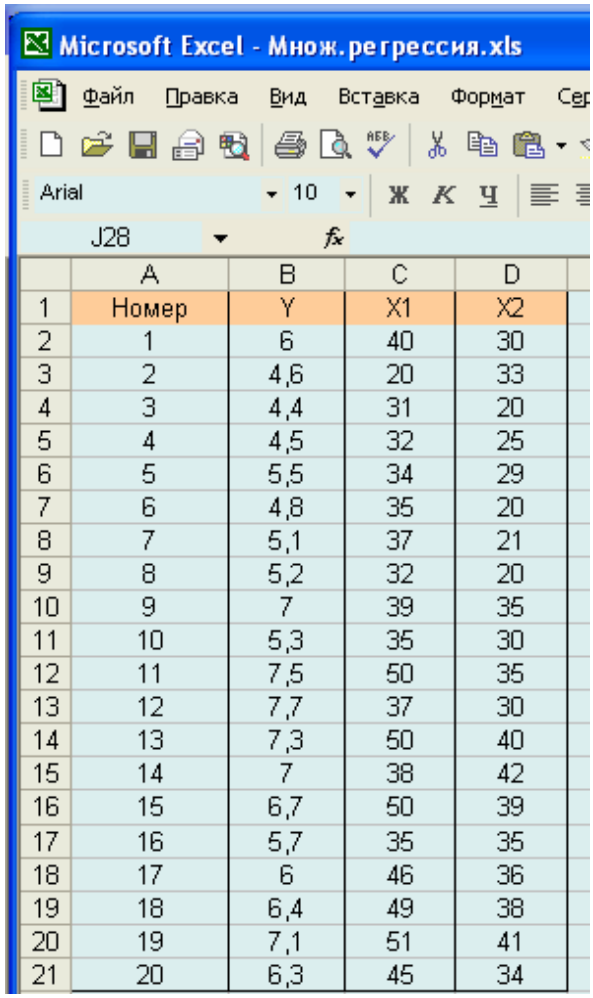
Исходные данные об эффективности продажи молока

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

3.1. 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 сколько-нибудь заметную тенденцию. Постараемся обеспечить решение задачи с использованием компьютерных программ в режиме Windows.

Запускаем Excel и воспроизводим в табличной форме имеющиеся исходные результаты (табл.7). В данном случае все экспериментальные данные (по каждой позиции) представляем в виде самостоятельных колонок (рис.14). 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.14. 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:

– в главном меню выбираем последовательно пункты **Сервис/Анализ данных/Описательная статистика**, после чего щелкаем по кнопке **ОК**;

– заполняем диалоговое окно для ввода данных и параметров вывода.

Чтобы получить их, проделаем следующие манипуляции (рис.15):

а) 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 ;

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

в) откроем флажок для **Метки**, показывающий, что первая строка содержит название столбца;

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

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

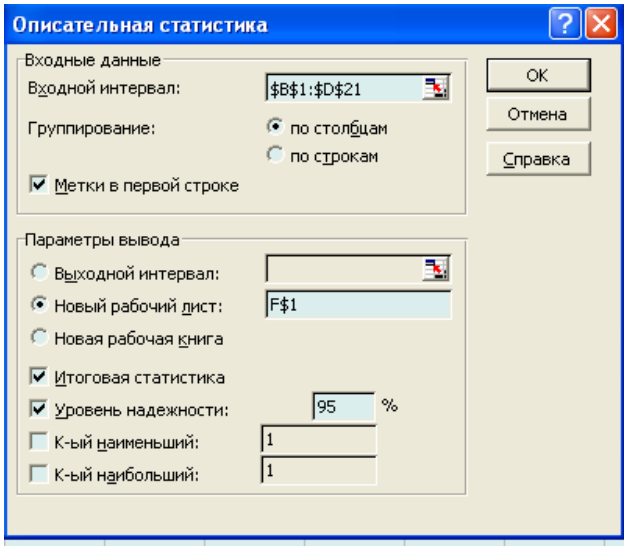


Рис.15. Диалоговое окно ввода параметров **Описательная статистика**

Полученные результаты статистического расчета показаны на рис.16 в виде соответствующего листа Excel.

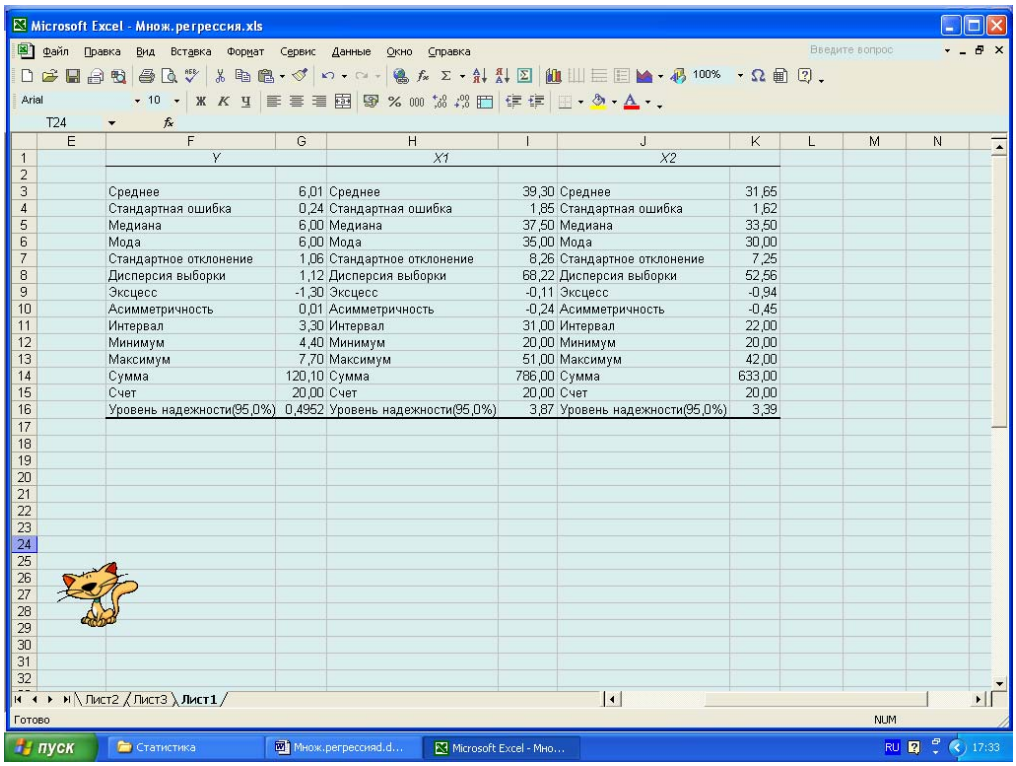
Из представленного комплекта статистических показателей выберем те, которые нам потребуются для последующего анализа – среднее арифметическое и стандартное отклонение (среднеквадратичное отклонение) S_n.

В табл.8 приведены названные статистические показатели для функции y и обеих переменных x_1 и x_2. Отметим, что для функции y ее среднее арифметическое $\bar{y}$ составляет 6,01, а стандартное отклонение (среднеквадратичное отклонение) S_n равно 1,06.

Таблица 8

Статистические показатели для функции y и переменных x_1 и x_2

Показатели	y	x_1	x_2
Среднее	6,01	39,3	31,65
Стандартное отклонение S_n	1,06	8,26	7,25



	E	F	G	H	I	J	K	L	M	N
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			

Рис.16. Лист Excel с результатами расчета статистических показателей

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 ее данные «сидят» во втором столбце B2:B21, а для переменных x_1 и x_2 – в третьем и четвертом, т.е. в диапазоне ячеек 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.17.

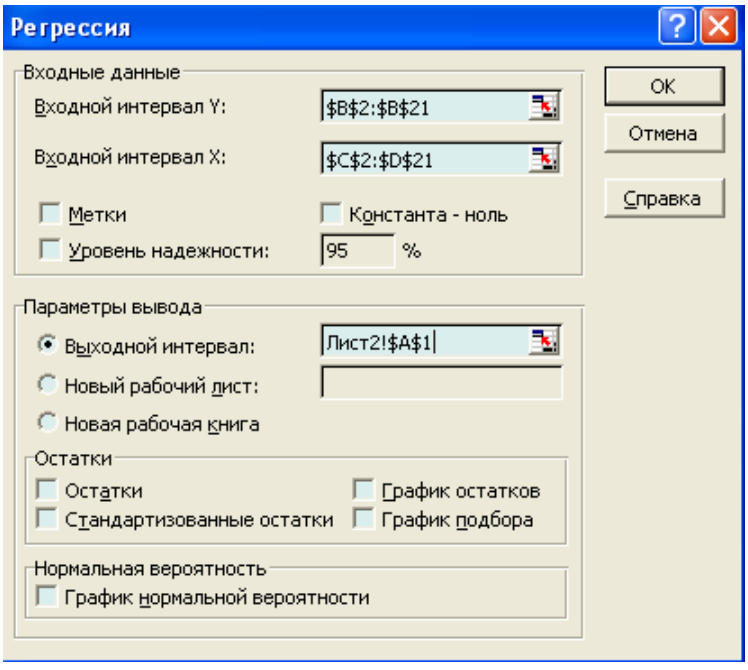


Рис.17. Диалоговое окно ввода параметров ***Регрессия***

Старательный Excel выдает, как мы убеждаемся, весьма богатый набор разнообразных статистических материалов (рис.18). Выберем, однако, из них такие, которые нам потребуются для последующего анализа.

Для этого создадим табл.9, 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*.

Таблица 9

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*: $S_{\bar{y}} = 0{,}65$; *R*-квадрат = 0,67; *R*-квадрат (нормир.) = 0,63. Таким образом, для рассматриваемого примера уравнение регрессии (или уравнение прогнозирования) будет иметь следующий вид:

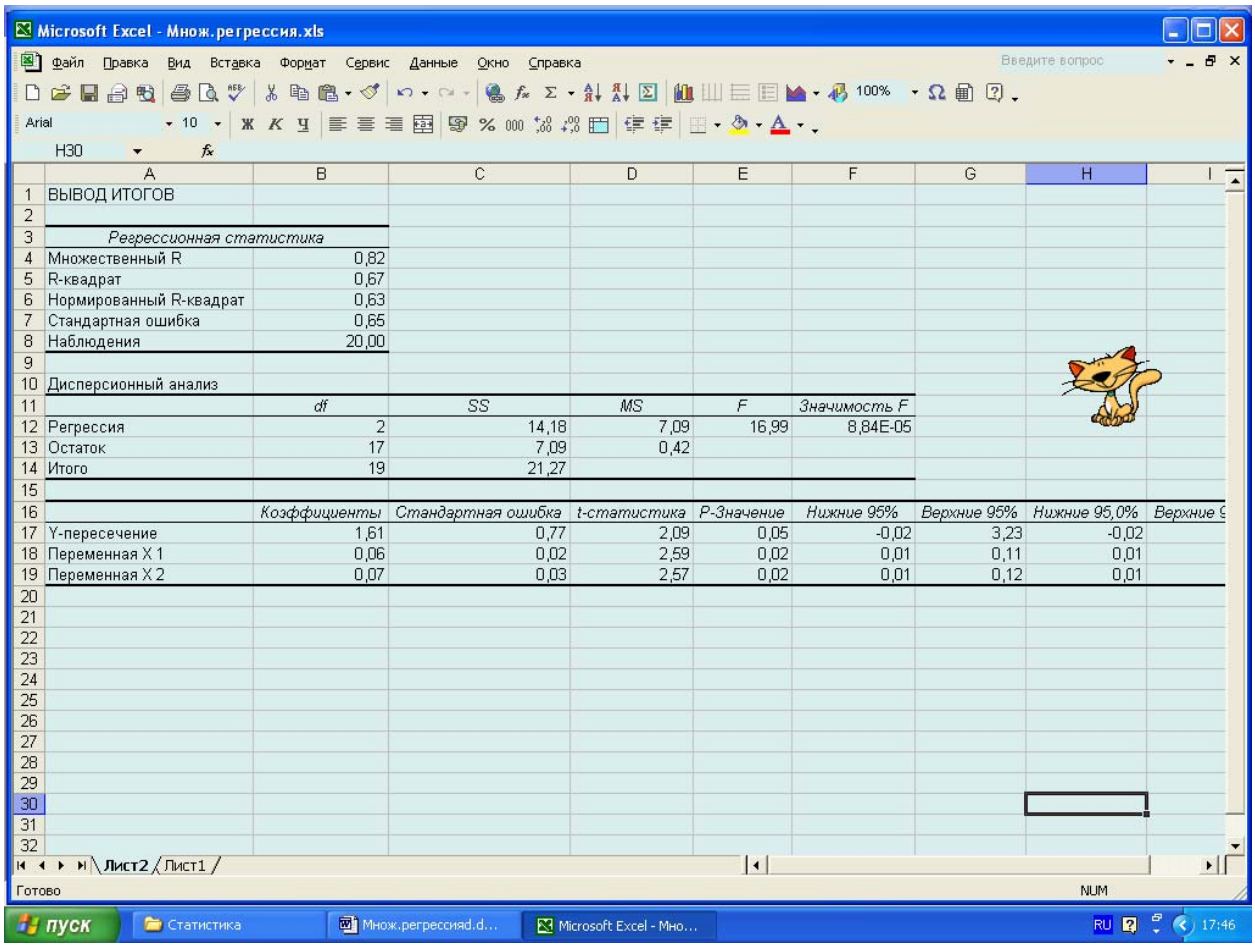
$\hat{y}$ (*объем продажи молока, л/день*) $= b_0 + b_1x_1 + b_2x_2 =$

$= 1{,}61 + 0{,}06$ (*доля среди покупателей бабушек с внучками*, *%*) +

$+ 0{,}07$ (*относительный вклад участия в торговле Шарика*, *%*).

Запишем полученное уравнение в окончательной редакции:

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



Microsoft Excel - Множ.регрессия.xls

Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

	A	B	C	D	E	F	G	H	I
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%*	*Верхние 9*
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									
21									
22									
23									
24									
25									
26									
27									
28									
29									
30									
31									
32									

Лист2 Лист1

Готово NUM

пуск Статистика Множ.регрессияd... Microsoft Excel - Мно... RU 17:46

Рис.18. Лист Excel с результатами расчета статистических показателей регрессии

Теперь займемся статистическим анализом этого уравнения регрессии.

3.2. Интерпретация коэффициентов регрессии

Свободный член (сдвиг) b_0, равный 1,61, формально надлежит понимать следующим образом: объем продажи молока котом Матроскиным, когда среди покупателей отсутствуют 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 помогает ежедневно). Поэтому сдвиг b_0 следует обсуждать как вспомогательную величину, необходимую для получения оптимальных прогнозов, и не истолковывать ее 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. Анализируя коэффициент b_2, можно заметить, что увеличение относительного участия Шарика на 1 % приводит также к повышению продажи, этот прирост составляет почти такую же величину – 0,07 л/день.

Еще 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. Например, 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.

3.3. 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_k.

Рассмотрим оба подхода более подробно.

1. Результаты статистического расчета показывают, что стандартная ошибка для функции составляет 0,65. Этот результат применительно к наше-

му примеру следует рассматривать следующим образом: фактическая величина объема продаж молока отличается от прогнозируемых показателей не более чем на 0,65 л/день. Однако ценность этого показателя невелика, если не надежность этого утверждения. При условии сохранения нормального распределения можно полагать, что примерно 2/3 фактических 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.8), которое было рассчитано для одной переменной, а именно: сами текущие значения y_iи величина среднего арифметического $\bar{y}$, которое равно 6,01. Легко видеть, что $S_{\bar{y}} < S_n$; следовательно, ошибки прогнозирования, как правило, оказываются меньшими, если использовать уравнение регрессии (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 высококвалифицированной помощи 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.17 славный Excel его подает как *R*-квадрат) равен 0,67, что составляет 67 %. Этот результат следует толковать так: все исследуемые воздействующие факторы (влияние особой категории покупателей и коммерческий талант 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, поэтому можно полагать, что именно эти два фактора в данном конкретном случае оказывают наиболее значительное влияние на $\bar{y}$.

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

Теперь нам надлежит определить статистическую значимость полученного аналитического выражения.

3.4. Проверка значимости модели

При проверке значимости модели принято придерживаться следующей последовательности действий:

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_k.

3.4.1. 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_k объяснять значимую часть колебания функции y. Этот тест используется как своеобразные «входные ворота» в статистический вывод: если результат теста значим, то связь существует, значит приступать к ее исследованию и объяснению. Если проверка указывает на незначимость связи, 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 какой-либо случайности нам не удалось ее доказать на основании тех опытных данных, которые были в нашем распоряжении.

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

С другой стороны, *альтернативная гипотеза F*-теста говорит о том, что между параметром y и переменными x_k существует определенная прогнозирующая взаимосвязь. Следовательно, параметр y уже не является чисто случайной величиной и должен зависеть хотя бы от одной из переменных x_k. Тем самым альтернативная гипотеза настаивает на том, что

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 $F_{\text{табл}}$ для соответствующего уровня значимости. Если выполняется неравенство $F_{\text{расч}} < F_{\text{табл}}$, 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. В противном случае 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 трактовать как незначимый (для 95-процентной вероятности). В том случае, когда величина *p* оказывается меньше 0,05, то вывод такой: это значимое уравнение с вероятностью 95%. Если же $p < 0{,}01$, то полученный результат является высоко значимым, (степень риска ошибиться в нашем утверждении оказывается меньше 1 %, т.е. степень надежности составляет 99 %)

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

В этом случае имеющуюся расчетную величину $R^2_{\text{расч}}$ (это то, что нам выдал Excel, см. рис.18) необходимо сравнить с табличными (критическими) значениями $R^2_{\text{крит}}$ для соответствующего уровня значимости (повторим еще раз, обычно это 0,05). Если окажется, что $R^2_{\text{расч}} > R^2_{\text{крит}}$, 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 $F_{\text{расч}}$, равную 16,99 (см. лист Excel на рис.18). С учетом сделанных замечаний (стр.36) для анализа уравнения будем пользоваться величиной $F_{\text{расч}}$, обратной представленной Excel. Она составит 1:16,99 = 0,06. Отыщем по эталонной таблице (*прил.3*) критическую величину $F_{\text{крит}}$ при условии, что для числителя степень свободы $f_1 = k$, т.е. составит 2 (число воздействующих факторов равно 2), а для знаменателя $f_2 = n - k - 1 = 20 - 2 - - 1 = 17$. Тогда будем иметь следующие значения для $F_{\text{крит}}$: 3,6 (для $\alpha = 0{,}05$), 6,2 ($\alpha = 0{,}01$) и 10,5 ($\alpha = 0{,}001$). Понятно, что для всех рассмотренных вероятностей выполняется соотношение $F_{\text{расч}} < F_{\text{крит}}$, 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.18 , 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 $\alpha = 0{,}000$. Это говорит о том, что действительно обнаруживается устойчивая зависимость рассматриваемой функции *y* (величины продажи молока) от воздействующих факторов x_1 и x_2, т.е. объем 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 $R^2_{\text{расч}}$ 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 $R^2_{\text{крит}} = 0{,}297$ (*прил.4*). Поскольку выполняется соотношение $R^2_{\text{расч}} > R^2_{\text{крит}}$, 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 $R^2_{\text{крит}}$ для α=0,01 (высокая значимость) и α = = 0,001 (высшая степень значимости). В этом случае $R^2_{\text{крит}}$ составляет соответственно 0,384 и 0,517, что, как видно, все равно остается меньше расчетного показателя $R^2_{\text{расч}}$, 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 способов анализа 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.

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

фициентов регрессии b_1 и b_2. Для этой цели используется так называемый *t*-тест.

3.4.2. Проверка на адекватность коэффициентов регрессии

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 *t-критерий* по абсолютной величине 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 всего вышесказанного 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). 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: $\hat{y} = 1{,}61 + 0{,}06\, x_1 + 0{,}07\, x_2$.

3.5. Сравнительная оценка степени влияния факторов

При анализе полученного уравнения множественной регрессии закономерно встает вопрос, а какой фактор x_k 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_k 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 меньший по величине коэффициент на деле может оказаться важнее большего. Короче говоря, в данной ситуации мы сталкиваемся с классической проблемой «попытки сравнения кита и слона». *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 стандартное отклонение S_n (для наших x-переменных обозначим его как S_{xk}) и деления полученного произведения на S_y. Это означает, что каждый стандартизованный коэффициент регрессии измеряется как величина $b_i\ S_{xk}\ /\ S_y$.Применительно к нашему примеру получим следующие результаты (табл.10).

Таблица 10

Стандартизованные коэффициенты регрессии

Статистические характиристики	Объем продажи	Бабушки с внучками	Помощь Шарика
Стандартные отклонения	$S_y = 1{,}06$	$S_{X1} = 8{,}26$	$S_{X2} = 7{,}25$
Коэффициенты регрессии	—	$b_1 = 0{,}06$	$b_2 = 0{,}07$
Стандартизованные коэффициенты регрессии	—	$b_1 S_{X1} / S_y =$ $0{,}06 \times 8{,}26/1{,}06=0{,}47$	$b_2 S_{X2}/ S_y =$ $=0{,}07 \times 7{,}25/1{,}06=0{,}48$

После проделанных расчетов мы можем на объективном основании сопоставить 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. Также оказывает влияние и конкретная категория покупателей – бабушки с внучками. Вместе с тем для Матроскина остаются поводы для творческих размышлений: он явно не принял во внимание все факторы (вспомним про 33 %, приходящихся на неучтенные причины), поскольку решил ограничиться рассмотрением более понятных и очевидных воздействий на результативность своего молочного бизнеса.

4. Анализ «хи-квадрат»: поиск закономерностей для качественных данных

Когда не знаешь, что именно ты делаешь,
делай это все-таки 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 объектов выборки, попадающих 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 он рассчитывается как сумма квадратов разности этих частот, выраженная в долях частоты теоретической. Это утверждение можно записать следующим образом: $\chi^2 = \Sigma \frac{(f_Э - f_Т)^2}{f_Т}$.

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

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

Данный способ широко применяется в тех случаях, когда нужно определить, 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), по характеру аналогичны прошлым данным (опорным величинам). Это объясняется тем, что и те, и другие относятся к одной и той же генеральной совокупности, но 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 ожидаемой* частотах анализируемого события определяется собственно показатель *хи-квадрат. Расчетное* значение *хи-квадрат* затем сравнивают с *критическим* (табличным) показателем для соответствующего числа степеней свободы, который 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 $\chi^2_{\text{расч}} < \chi^2_{\text{крит}}$, то наблюдаемые значения незначительно 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 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 ее 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:

$$\chi^2 = \Sigma \frac{(f_{\text{Э}} - f_{\text{Т}})^2}{f_{\text{Т}}}.$$

Степень свободы f рассчитывается так:

$$f = k - 1,$$

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 конкретному анализу критерия соответствия частот и, самое главное, выясним, как такой расчет можно выполнить с использованием компьютерной программы Excel.

Рассмотрим следующий пример.

Среди студентов металлургического факультета, сдававших на первом курсе в летнюю сессию экзамен по математике, был проведен опрос с целью выяснения того, какие факторы влияют на получение неудовлетворительной оценки. Число опрошенных студентов составляло 50 человек.

Наиболее часто упомянутыми причинами были 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.11 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 из данных таблицы, по количественным показателям все анализируемые причины 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».

Таблица 11

Итоговые данные о причинах получения неудовлетворительной оценки по математике за анализируемую сессию и сессии прошлых лет

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 ли велика 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 $\chi^2_{\text{расч}} < \chi^2_{\text{крит}}$, то с заданной вероятностью можно будет говорить о незначимости различия между анализируемыми результатами.

В табл.12 укажем частотные величины для обеих информационных позиций – текущие данные («Наблюдение») и сведения за прошлые годы («Ожидание»). Расчет частот для графы «Ожидание» (т.е. ожидаемые частоты) 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.

Таблица 12

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 расторопного Excel.

Примечание. 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 (см. *прил.5*). Однако в нашем случае используем возможности Excel, поскольку подобную услугу он способен оказать совершенно элементарно.

Откроем лист Excel и составим нашу таблицу с имеющими данными (рис. 19). 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).

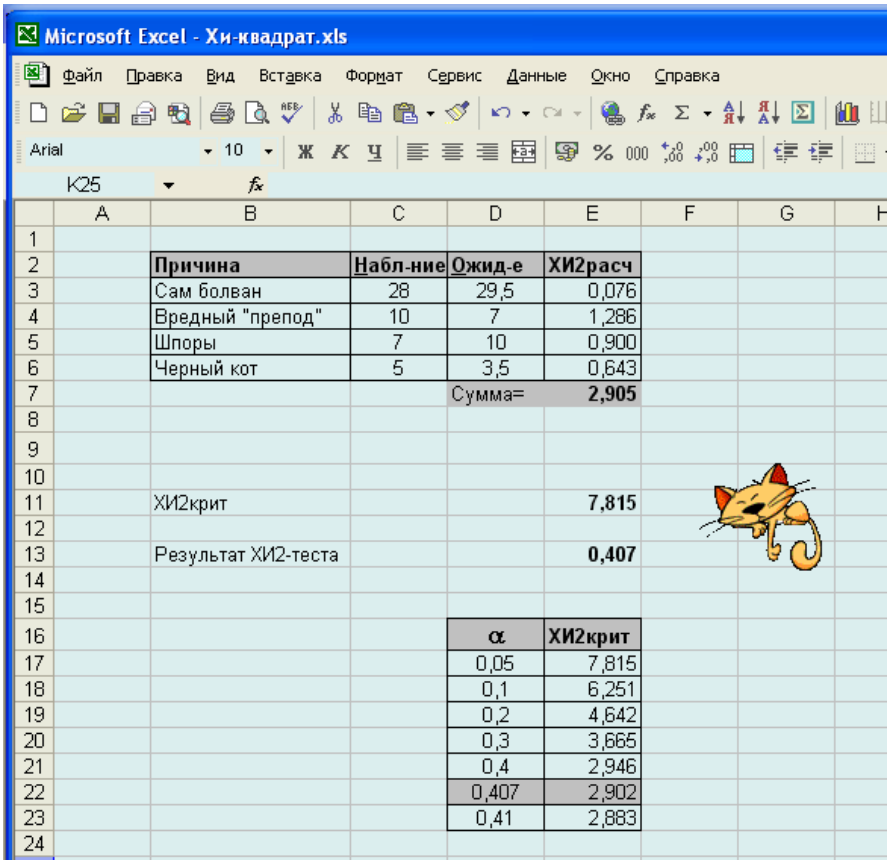
Расчет проведем по уже известной формуле, запись которой представлена в виде:

$$\text{ХИ2расч} = \Sigma(f_{\text{Э}} - f_{\text{Т}})\text{^}2/f_{\text{Т}},$$

где $f_{\text{Э}}$ и $f_{\text{Т}}$ – соответственно экспериментальные (наблюдаемые) и теоретические (ожидаемые) значения частот.

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 $\chi^2_{\text{расч}}$.



Microsoft Excel - Хи-квадрат.xls

Файл Правка Вид Вставка Формат Сервис Данные Окно Справка

Arial 10

K25

	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.19. Фрагмент рабочего листа Excel с исходными данными и результатами анализа *хи-квадрат*

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**.

На экране появится диалоговое окно для ввода параметров, необходимых для вычисления критического (табличного) значения *хи-квадрата* (рис.20). В первом текстовом поле ввода (**Вероятность**) укажем выбранную величину уровня значимости α. Примем традиционный показатель степени риска, равный 0,05.

Во втором поле ввода (**Степени свободы**) запишем число степеней свободы. В нашем примере фигурируют четыре компонента (причины "неудов"), поэтому 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).

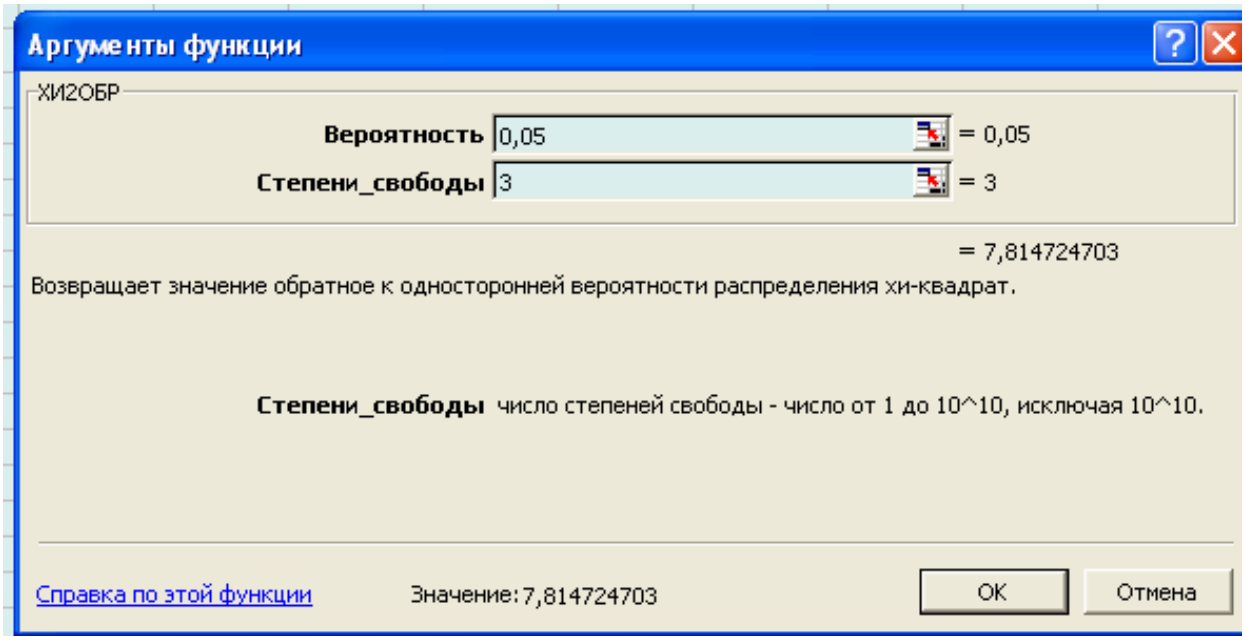


Рис.20. Диалоговое окно ввода параметров
для определения критического(табличного) значения *хи-квадрат*

Вот с этим-то числом нам и нужно теперь сравнить расчетное значение $\chi^2_{расч}$. Поскольку выполняется неравенство $\chi^2_{расч} < \chi^2_{крит}$ (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.

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**.

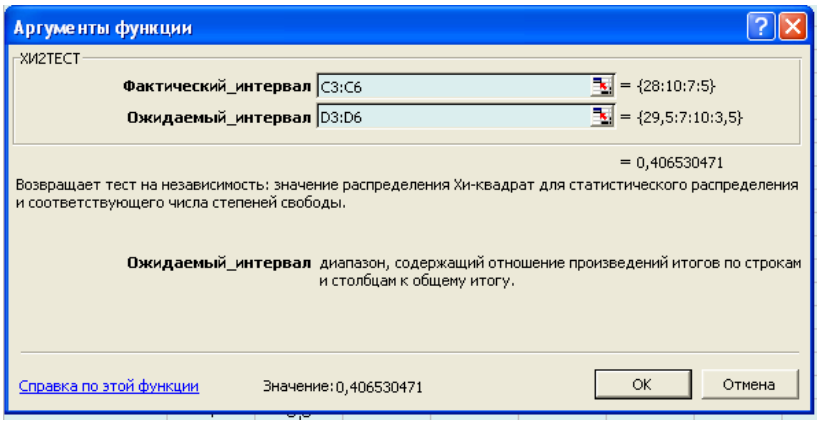


Рис.21. Диалоговое окно ввода параметров
для определения расчетногозначения *хи-квадрат*

В появившемся диалоговом окне (рис.21) 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.

Кстати, после введения интервальных ячеек справа от каждого поля ввода в скобках будут перечислены те табличные значения, которые содержались в соответствующих столбцах (рис.21). 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 обсудить полученные данные.

Указанное число показывает: 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 $\alpha = 0{,}05$ (т.е. для 5-процентной степени риска). Теперь попытаемся выяснить, при каких же условиях можно отважиться на утверждение, что экзаменационные данные нынешней и прошлых сессий (по характеру рассматриваемых факторов влияния на их итоги) все-таки разнятся. Иными словами, определим, когда можно считать, что эти данные являются извлечением не из одной и той же генеральной совокупности, а принадлежат к 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.19) в виде списка приведены полученные значения $\chi^2_{\text{крит}}$ для α, равного соответственно 0,05; 0,1; 0,2 и т.д. Закончим расчет и для случая $\alpha = 0{,}407$ и 0,41. Почему надо учесть именно эти числа, обусловлено следующим.

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 $\alpha = 0{,}41$ уже можно определенно говорить, что условие $\chi^2_{\text{расч}} > \chi^2_{\text{крит}}$ (2,905* > 2,883) выполняется. Поэтому допустимо утверждение, что обе рассматриваемые совокупности являются различными.

2. Теперь дадим оценку только что сделанному заявлению. Прелесть статистики состоит в том, что она любое утверждение дает с определенной гарантией надежности, т.е. страхуется от проявления возможных случайностей (погрешностей). Совершенно недостаточно высказать какое-то соображение. 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 $\alpha = 0{,}05$, а еще лучше 0,01), величина порогового (критического) значения χ^2 имеет очевидную тенденцию к возрастанию. А это означает, при разумном объеме единиц наблюдения (в данном случае это студенты, большие знатоки математической науки) мы можем говорить лишь о незначимости рассматриваемых итогов. Чтобы все-таки обнаружить подобное возможное различие, следовало бы провести более масштабное по охвату обследование. Однако можно утешиться тем обстоятельством, что проделать всю эту процедуру весьма проблематично вследствие недостаточного числа (смеем надеяться!) физически наличествующих двоечников.

*Числа 2,902 и 2,905 - это фактически одно и то же, различие обусловлено некоторым искажением при выполнении операции округления

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 и методической комиссии факультета нет повода для беспокойства.

4.2. 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:

$$K_{\text{П}} = \sqrt{\frac{\chi^2}{n + \chi^2}},$$

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:

$$K_{\text{ч}} = \sqrt{\frac{\chi^2}{n\sqrt{(k_1 - 1)(k_2 - 1)}}},$$

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 (объем выборки).

Попробуем выполнить такие расчеты для нашего примера.

$$K_{\text{П}} = \sqrt{\frac{\chi^2}{n+\chi^2}} = \sqrt{\frac{2{,}905}{50+2{,}905}} = 0{,}234;$$

$$K_{\text{ч}} = \sqrt{\frac{\chi^2}{n\sqrt{(k_1-1)(k_2-1)}}} = \sqrt{\frac{2{,}905}{50\times\sqrt{(4-1)(2-1)}}} = 0{,}184.$$

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 меньшие 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.

4.3. Проверка наличия взаимосвязи между двумя качественными переменными (критерий «хи-квадрат» независимости)

Возможны ситуации, когда имеются две качественные переменные, т.е. набор экспериментальных данных представляет собой *двумерные качественные* данные. 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 Пупкин, 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 Пупкина 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 следующим образом:

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 (теоретических) частот*. Для их расчета частоту одной категории (результат эксперимента) следует умножить на частоту другой категории (также экспериментальный показатель) и полученное произведение поделить на общий объем выборки *n*:

$$\text{Ожидаемая частота } f_{\text{ОЖ(Т)}} = \frac{|\text{Частота категории } f_{\text{Э1}}| \text{ для одной переменной} \times |\text{Частота категории } f_{\text{Э2}}| \text{ для другой переменной}}{\text{Общий объем выборки } n},$$

или более компактно, в символьной форме: $f_{\text{ОЖ(Т)}} = \dfrac{f_{\text{Э1}} \times f_{\text{Э2}}}{n}$.

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-квадрат2* 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.

Давайте 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.13), в которой для каждого ви-

да 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 – определить 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 – традиционным способом («вручную») и компьютерным.

Таблица 13

Результаты опроса о перспективах молочного бизнеса

Вид молочной продукции	Частота предпочтений	
	Практичный покупатель	Импульсивный покупатель
Молоко	38	15
Сметана	24	31
Творог	18	27

Для этого дополним таблицу с исходными данными итоговой строкой и дополнительным «суммирующим» столбцом, заполним их, выполнив несложные расчеты (табл.14).

Таблица 14

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

Чисто визуально трудно ответить, есть ли взаимосвязь между этими признаками: разными категориями покупателей и видами молочной продукции. Поэтому необходимо дать анализ распределения частот в таблице по строкам и графам.

Будем исходить из следующего положения. 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: $f_{11} = \frac{53 \times 80}{153} = 27{,}7$; $f_{12} = \frac{53 \times 73}{153} = 25{,}3$.

Для второй строки, т.е. для категории *сметана*, эти показатели имеют уже такие значения:

$$f_{21} = \frac{55 \times 80}{153} = 28{,}8; \; f_{22} = \frac{55 \times 73}{153} = 26{,}2.$$

Для третьей строки (категория *творог*):

$$f_{31} = \frac{45 \times 80}{153} = 23{,}5; \; f_{32} = \frac{45 \times 73}{153} = 21{,}5.$$

Полученные результаты (вычисленные значения частот) поместим в табл.15.

Таблица 15

Данные о перспективах молочного бизнеса с учетом ожидаемых частот

Вид молочной продукции	Ожидаемая частота предпочтений		Итого
	Практичный покупатель	Импульсивный покупатель	
Молоко	27,7	25,3	53
Сметана	28,8	26,2	55
Творог	23,5	21,5	45
Итого:	80	73	153

Расчетное значение критерия *хи-квадрат* определим по формуле:

$$\chi^2 = \sum_{i=1}^{k_1}\sum_{j=1}^{k_2}\frac{(f_{ij} - f_{ij}^*)^2}{f_{ij}^*},$$

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:

$$\chi^2_{\text{расч}} = \frac{(38-27{,}7)^2}{27{,}7} + \frac{(15-25{,}3)^2}{25{,}3} + \frac{(24-28{,}8)^2}{28{,}8} + \frac{(31-26{,}2)^2}{26{,}2} +$$

$$+ \frac{(18-23{,}5)^2}{23{,}5} + \frac{(27-21{,}5)^2}{21{,}5} = 12{,}4$$

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 двум, т.е. $f = (3 - 1)(2 - 1) = 2$*. При

* В данном случае и частота и степень свободы обозначены одним и тем же буквенным символом *f*.

α = 0,05 табличное значение $\chi^2_{\text{табл}}$ при f = 2 составляет 5,991 (*прил.5*), а для α = 0,01 соответственно 9,210. Поскольку $\chi^2_{\text{расч}} > \chi^2_{\text{табл}}$, то с уверенностью на 95 % (даже на 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 расторопный Excel.

Прежде всего, перенесем данные табл.13 и 14 в рабочий лист Excel (рис.22). 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 значений уровня значимости – 0,05 и 0,01. После этого приступим собственно к работе в компьютерном варианте.

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 – 2.

После исполнения всех манипуляций и необходимых округлений в ячейках В22 и С22 будут содержаться следующие результаты: 5,991 и 9,210.

Затем произведем необходимые подсчеты ожидаемых частот. Используем уже знакомое выражение:

$$f_{\text{ОЖ(Т)}} = \frac{f_{\text{Э1}} \times f_{\text{Э2}}}{n}.$$

Здесь поступим следующим образом. Вычисленные значения будем помещать в диапазоне ячеек В11:С13. Запишем формулу вычисления ожидаемых частот, которую затем скопируем 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).

Чтобы получить остальные значения ожидаемых частот, проделаем следующее. Выделим ячейку B11, в которой сидит наш первый вычисленный показатель, поднесем курсор к нижнему правому углу и, как только появится маленький черный крестик, 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.14 с ожидаемыми частотами. Что-то очень знакомое! С учетом необходимых округлений они почти полные копии друг друга.

Теперь мы наглядно представляем, насколько легко Excel справляется с расчетами, над которыми нам перед этим (вспомним ручной счет) пришлось изрядно потрудиться.

Анализ *хи-квадрат* выполним с помощью функции **ХИ2ТЕСТ**. Действием уже привычным образом, используя следующие команды: **Мастер функций/ Статистические / Хи2тест**.

Ячейку B32 выделим для **ХИ2ТЕСТ**.

При заполнении диалогового окна в текстовом поле *фактического интервала* укажем адрес ячеек B6:C8, в которых находятся экспериментальные данные по частотам (табл.13). Соответственно в текстовом поле *ожидаемого интервала* укажем диапазон B16:C18, содержимое которого отражает теоретические значения частот (табл.14).

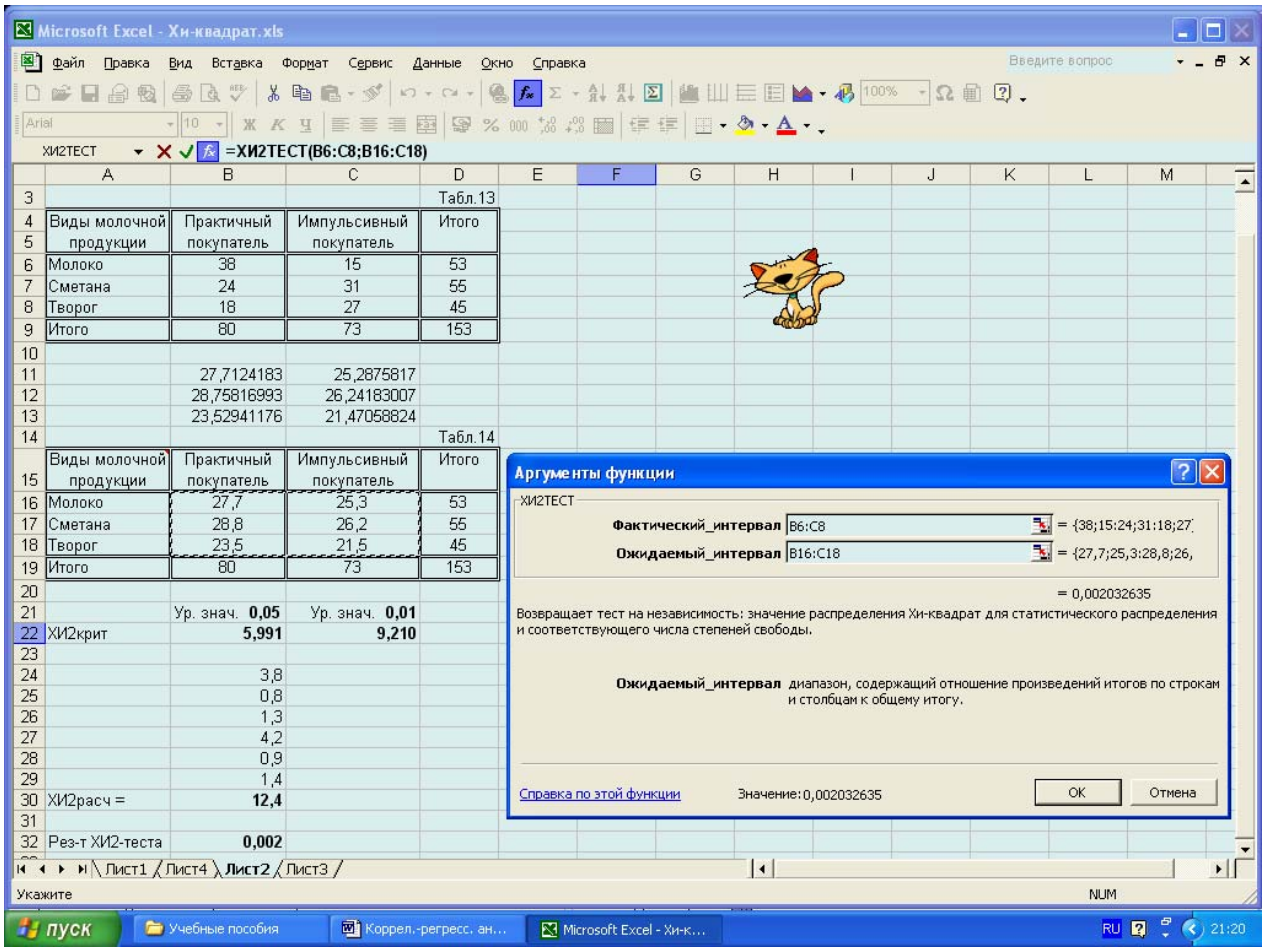


Рис.22. Лист Excel с результатами расчета критерия *хи-квадрат*

В окончательном виде в ячейке B32 будет находиться следующий показатель, а именно: 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 %−ной вероятности (даже 99-процентной) установленные значения α (0,05 и 0,01) превышают 0,002. Это говорит о высокой степени значимости, следовательно, указанные качественные переменные являются зависимыми друг от друга.

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 $\chi^2 = \frac{(f_Э - f_Т)^2}{f_Т}$ для каждой комбинации *наблюдаемых* (экспериментальных) $f_Э$ и *ожидаемых* (теоретических) *частот* $f_Т$ вычислим текущие значения χ^2, а затем их просуммируем. Результат приведен в виде списка на рис.22 (диапазон ячеек B24:B29) он, как и в случае ручного счета, равен 12,4 (ячейка B30). Дальше знакомые процедуры – сопоставление значений $\chi^2_{\text{расч}}$ (12,4), с одной стороны, и $\chi^2_{\text{табл}}$ (7,815 и 11,345), с другой, указывает на то, что анализируемые качественные переменные не являются независимыми (мы это утверждаем с риском ошибиться на 5 и даже 1 %). 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, коту Матроскину, как мы и полагали по итогам ручного счета, надлежит внимательно продумать свою дальнейшую коммерческую стратегию – продаваемая продукция существенного зависит от того, кто ее покупает. Причем наиболее заметно это проявляется в торговле 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, предпочитая сметану. Такого рода соображения можно высказать на основании выполненного анализа.

Приложения

Статистико-математические таблицы

Приложение 1

Критические значения корреляции $\tau_{крит}$ 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 2

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 3

Значения *F*-критерия для уровня значимости $\alpha = 0{,}05$ и числа степеней свободы *f*

Знаменатель: степени свободы *f*	**Числитель: степени 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 4

Критические значения R^2 для уровня значимости α, числа переменных (аргументов) *x* и количества опытов *n*

Уровень значимости		α = 0,1			α = 0,05			α = 0,01		
Число переменных *x*		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 5

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

Приложение 6

Значения коэффициента Стьюдента (*t*-критерия) для уровня значимости α и числа 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

Библиографический список

1. Бараз, В.Р. Применение программы Excel для статистических расчетов в материаловедении : учебное пособие. – Екатеринбург : ГОУ ВПО «УГТУ-УПИ», 2003. – 46 с.
2. Сигал, Э. Практическая бизнес-статистика. – М. : издательский дом «Вильямс», 2002. – 1056 с.
3. Годин, А.М. Статистика : учебник. М. : издательско-торговая корпорация «Дашков и К°», 2002. – 368 с.
4. Хайкин, Б.Е. Построение аппроксимационных математических моделей в условиях обработки металлов давлением : учебное пособие. – Свердловск : УПИ, 1991. – 101 с.
5. Макарова , Н. В., Трофимец В.Я. Статистика в Excel : учебное пособие. – М. : Финансы и статистика, 2002. – 192 с.
6. Информатика. Базовый курс / Под ред. С.В.Симоновича. – СПб. : Питер, 2001. – 640 с.
7. Нельсон , С. Анализ данных в Excel для «чайников». – М. : издательский дом «Вильямс», 2002. – 302 с.

Учебное пособие

Бараз Владислав Рувимович

Корреляционно-регрессионный анализ связи показателей коммерческой деятельности с использованием программы Excel

Редактор *Е.А.Сенкевич*
Компьютерная верстка *Авторская*

ИД № 06263 от 12.11.2001 г.

Подписано в печать 01.09.05		Формат 60x84 1/16
Бумага писчая	Плоская печать	Усл. п. л_____
Уч.-изд. л. 3,4	Тираж _____	Заказ _____

Редакционно-издательский отдел ГОУ ВПО УГТУ-УПИ
Ризография НИЧ ГОУ ВПО УГТУ-УПИ
620002, г. Екатеринбург, ул. Мира, 19