Н.В.МАКАРОВА, В.Я.ТРОФИМЕЦ

СТАТИСТИКА В Excel

Н.В.МАКАРОВА, В.Я.ТРОФИМЕЦ

СТАТИСТИКА в Excel

Рекомендовано
Учебно-методическим объединением вузов РФ
по образованию в области прикладной информатики,
статистики и математических методов в экономике
в качестве учебного пособия для студентов,
обучающихся по специальности
061700 "Статистика" и другим специальностям

Москва
"Финансы и статистика"
2003

УДК 004.67:311(075.8)
ББК 65.051с51я73
М15

РЕЦЕНЗЕНТЫ:
кафедра экономической кибернетики
и экономико-математических методов
Санкт-Петербургского
государственного университета
экономики и финансов;
Л. Г. Батракова,
доктор экономических наук, профессор

Макарова Н. В., Трофимец В. Я.

М15 Статистика в Excel: Учеб. пособие. – М.: Финансы и статистика, 2003. – 386 с.: ил.

ISBN 5-279-02282-9

Рассматриваются функциональные возможности табличного процессора Excel для проведения статистического анализа данных на персональном компьютере. Описывается технология работы с программной надстройкой «Пакет анализа» и встроенными статистическими функциями. Приводится большое количество примеров по обработке экономической информации. Содержатся краткие сведения из теории статистики, помогающие читателю быстрее разобраться с существом реализованных в Excel статистических методов.

Для студентов, аспирантов, преподавателей, экономистов, инженерно-технических работников, занимающихся статистической обработкой данных.

М $\frac{0702000000 - 200}{010(01) - 2003}$ 357 – 2003

УДК 004.67:311(075.8)
ББК 65.051с51я73

ISBN 5-279-02282-9

ПРЕДИСЛОВИЕ

В современном обществе к статистическим методам проявляется повышенный интерес как к одному из важнейших аналитических инструментариев в сфере поддержки процессов принятия решений. Статистикой пользуются все – от политиков, желающих предсказать исход выборов, до предпринимателей, стремящихся оптимизировать прибыль при тех или иных вложениях капитала. Большим шагом вперед к развитию статистической науки послужило применение экономико-математических методов и использование компьютерной техники в анализе социально-экономических явлений.

Стандартные статистические методы обработки данных включены в состав электронных таблиц, таких, как Lotus 1-2-3, QuattroPro, Excel и др.; в математические пакеты общего назначения – Mathcad, Mathlab, Maple и т.д. Еще более мощными возможностями статистической обработки обладают специализированные пакеты, как отечественные – STADIA, МЕЗОЗАВР, СИГАМД, СТОД, САНИ, ОЛИМП:СтатЭксперт и др., так и зарубежные – STATGRAPHICS, SPSS, SAS, BMDP, STATISTICA и др.

Наибольшее распространение в деловой сфере получил табличный процессор Microsoft Excel, который, по данным еженедельника ComputerWeek, еще в конце 1995 г. использовали в своей деятельности более 60 % московских организаций, в том числе и для статистического анализа информации. За последние пять лет популярность Excel еще более возросла, что объясняется его органичной интеграцией в пакет Microsoft Office (начиная с Microsoft Excel 7.0 for Windows 95).

Для проведения статистической обработки информации табличный процессор Microsoft Excel включает в себя программную надстройку «Пакет анализа» и библиотеку из 78 статистических функций. В повседневной деятельности такого набора инструментов бывает, как правило, вполне достаточно для проведения

довольно полного и качественного статистического анализа информации. Если же пользователя не удовлетворяют подобные возможности Excel, тогда необходимо обратиться к мощным специализированном пакетам статистического анализа, в частности к пакету STATISTICA фирмы StatSoft.

В настоящее время существует разнообразная литература по работе с Excel различных версий, адресованная как новичкам, так и опытным пользователям. В этих изданиях приводятся достаточно подробные сведения о порядке и правилах работы с Excel, рассматриваются многочисленные примеры, даются практические рекомендации. Помимо этого, сама справочная система Excel является мощным путеводителем, способным оказать помощь в самых различных ситуациях. Вместе с тем, несмотря на такую сильную обучающую поддержку, по мнению разработчиков из Microsoft, «средний пользователь» применяет только 5 % функциональных возможностей, заложенных в Excel. Такую «неграмотность среднего пользователя» можно объяснить, по всей видимости, несколькими причинами.

Во-первых, Excel – очень мощный, достаточно универсальный табличный процессор, ориентированный на различные сферы деятельности, вследствие чего «проблемно-ориентированному среднему пользователю» просто нет необходимости обращаться к не интересующим его функциональным возможностям программы.

Во-вторых, значительная мощь Excel заключена в дополнительных программных надстройках и библиотеке аналитико-расчетных функций, которым в литературе уделяется, как правило, незначительное внимание (в лучшем случае дается конспективный обзор программных надстроек и приводится краткий глоссарий по функциям листа). На компьютерных курсах также нет времени подробно останавливаться на этих вопросах в силу их объемности и специфичности. Поэтому для «среднего пользователя» остается единственный выход – обратиться за помощью к справочной подсистеме программы, что, как замечено, он не очень-то любит делать. Но даже в том случае если пользователь окажется «продвинутым» и ему не составит труда познакомиться с содержанием соответствующих справочных разделов, то и здесь его могут поджидать уныние и разочарование, так как применение некоторых программных надстроек и большого числа функций Excel (за исключением тривиальных функций типа суммирования,

умножения, тригонометрических вычислений и т. п.) требует математической подготовки. Поэтому, чтобы грамотно и осознанно применить какую-либо надстройку или функцию, пользователю придется обратиться к литературе для выяснения физической сущности соответствующего метода или функции. Данное обстоятельство можно в полной мере отнести и к программной надстройке «Пакет анализа», и к большинству статистических функций Microsoft Excel.

Справедливости ради следует отметить, что в справочной подсистеме Excel приводится краткая информация и по надстройке «Пакет анализа», и по каждой статистической функции, нередко рассматриваются примеры их практического использования. Зачастую уже этого оказывается достаточно для грамотной работы с ними. Однако так обстоит дело не всегда. Часто появляется необходимость во вспомогательной информации, которая может быть рассредоточена по различным источникам. Требовать того, чтобы и она была включена в справочную подсистему, неразумно и бессмысленно, так как справочная система может возрасти до невообразимых размеров и при этом все равно не удовлетворит запросов всех пользователей в силу различия специфики их профессиональной деятельности и уровня образования.

Настоящее учебное пособие призвано помочь тем пользователям, которые используют или собираются использовать табличный процессор Excel для статистического анализа данных, – студентам, аспирантам, слушателям факультетов повышения квалификации, экономистам различного профиля.

Последовательность изложения материала в пособии соответствует порядку, принятому в большинстве учебников по статистике, вышедших в свет в издательстве «Финансы и статистика» (*см.* список литературы). Большое число примеров также заимствованы или перекликаются с примерами из данных учебников, поэтому, на наш взгляд, особенно эффективным может стать их совместное изучение.

Все примеры, рассмотренные в пособии, реализованы авторами в среде Microsoft Excel 97/2000. При апробировании этих примеров читателем возможны некоторые незначительные расхождения в получаемых результатах, что объясняется выбранным форматом соответствующих ячеек.

Главы книги имеют относительную законченность и могут изучаться лицами, знакомыми с общей теорией статистики, в любом порядке. В тех случаях, где все же рекомендуется предварительно ознакомиться с материалом предыдущих глав, сделаны соответствующие ссылки.

В приложении рассмотрены примеры комплексного использования надстройки «Пакет анализа» в ходе проведения статистического исследования.

Для быстрого поиска статистических функций в конце пособия приведен алфавитный указатель.

При изложении материала использованы конструкции, терминология и синтаксис табличного процессора Microsoft Excel 97/2000.

Остается выразить надежду, что настоящее учебное пособие поможет вам в полной мере оценить возможности Microsoft Excel в статистической обработке информации и станет незаменимым помощником в работе.

РАЗДЕЛ I

Методы описательной статистики

ГЛАВА 1
Общие сведения о надстройке «Пакет анализа» и статистических функциях MS Excel

1.1.
Первое знакомство с надстройкой «Пакет анализа»

1.1.1.
Установка надстройки «Пакет анализа»

При создании новой или открытии существующей книги Microsoft Excel появится окно активного рабочего листа. Для того чтобы отыскать команду вызова надстройки ***Пакет анализа***, необходимо воспользоваться меню ***Сервис*** (рис. 1.1).

Здесь возможны три ситуации, в которых нужно действовать следующим образом:

1. В меню ***Сервис*** присутствует команда **Анализ данных...** (рис. 1.1). Это идеальный случай – достаточно щелкнуть указателем мыши по данной команде, чтобы попасть в окно надстройки.

2. В меню ***Сервис*** отсутствует команда **Анализ данных...**. В этом случае необходимо в том же меню выполнить команду **Надстройки...**. Раскроется одноименное окно (рис. 1.2) со списком доступных надстроек. В этом списке нужно найти элемент *Пакет анализа*, поставить рядом с ним «галку» и щелкнуть по кнопке **OK**. После этого в меню ***Сервис*** появится команда **Анализ данных...**.

Эта ситуация наиболее типична, так как надстройка ***Пакет анализа*** инсталлируется при стандартной установке.

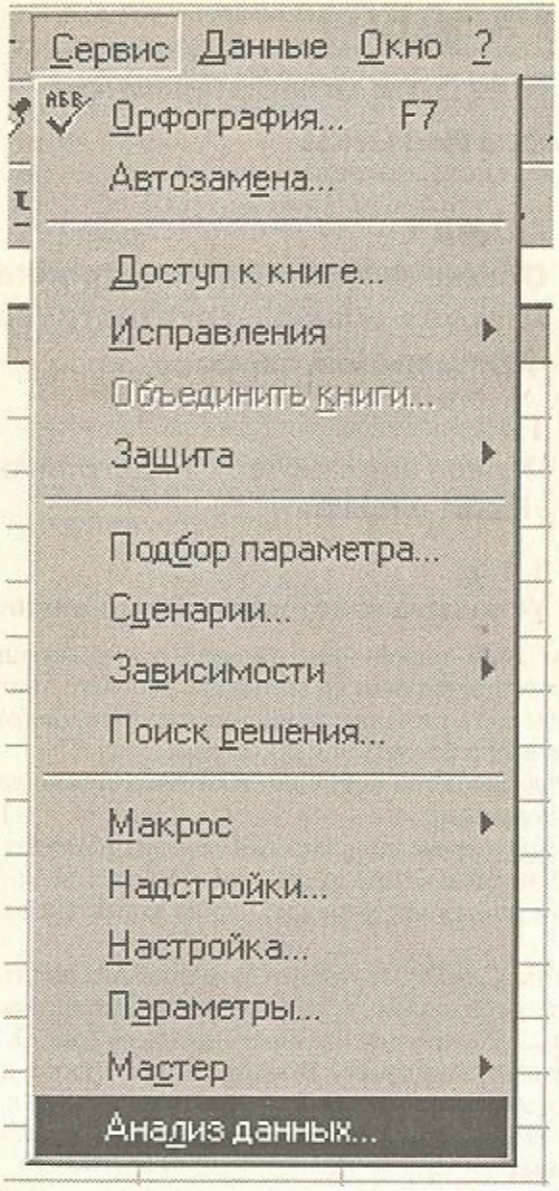

Рис. 1.1

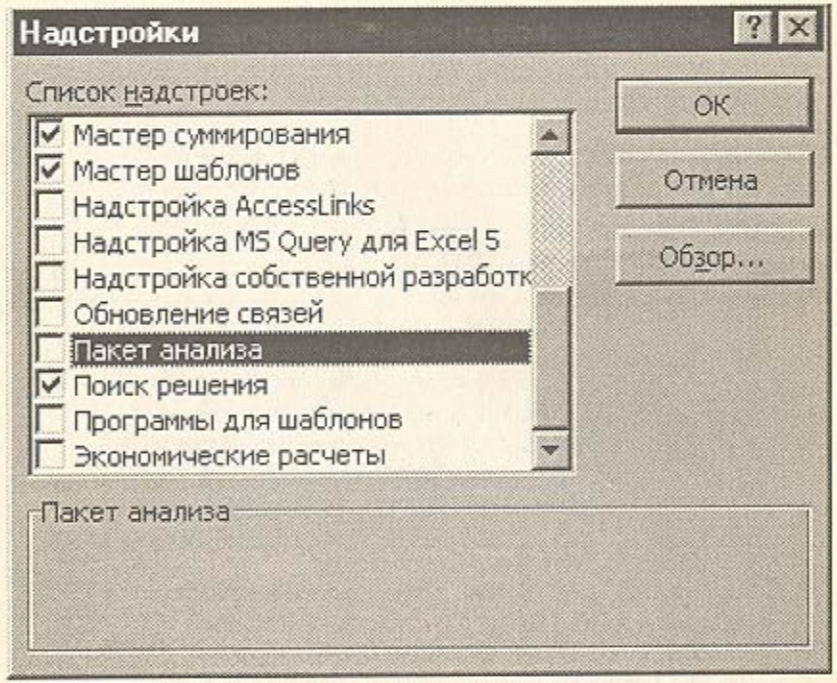

Рис. 1.2

3. В меню ***Сервис*** отсутствует команда **Анализ данных...**, а в списке окна **Надстройки** нет элемента *Пакет анализа*. Это самая неприятная ситуация, так как без установочного комплекта дискет или компакт-диска в этом случае не обойтись. Рассмотрим наиболее распространенный случай доустановки Excel с дистрибутивного компакт-диска Microsoft Office*.

После того как компакт-диск с пакетом Microsoft Office вставлен в CD-ROM, нужно перейти в папку ***Панель управления*** (один из возможных способов этого – выбрать в главном меню ***Пуск*** пункт **Настройка**, затем пункт **Панель управления**). В папке ***Панель управления*** дважды щелкните по значку *Установка и удаление программ*, после чего раскроется соответствующее окно (рис. 1.3).

* Начиная с версии 7.0, Excel входит в состав пакета Microsoft Office.

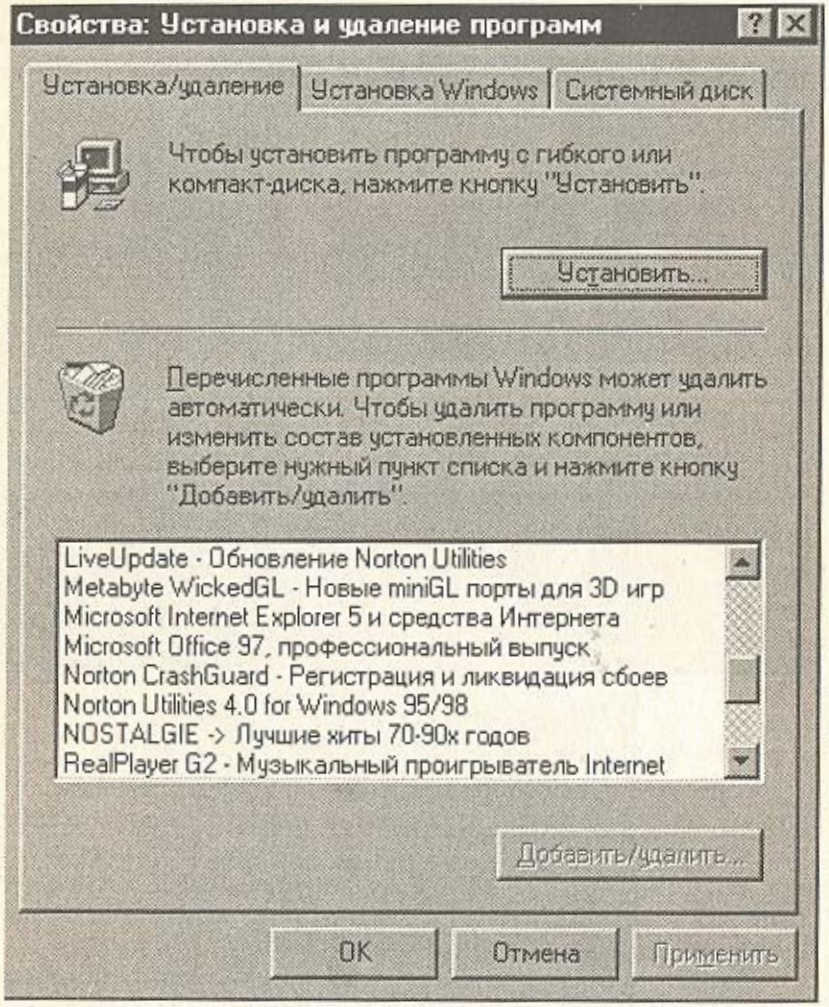

Рис. 1.3

Выберите в списке установленных программ элемент *Microsoft Office 97, профессиональный выпуск* и щелкните по кнопке **Добавить/удалить**, после чего отвечайте по умолчанию на сообщения, выводимые программой установки, пока не появится окно **Установка Microsoft Office 97**.

В дальнейшем порядок действий следующий:

1) в окне **Установка Microsoft Office 97** щелкните по кнопке **Добавить/удалить** – откроется окно **Microsoft Office 97-Сопровождение**;

2) в списке окна **Microsoft Office 97-Сопровождение** выберите элемент *Excel* и щелкните по кнопке **Состав...** – откроется окно **Microsoft Office 97-Microsoft Excel**;

3) в списке окна **Microsoft Office 97-Microsoft Excel** выберите элемент *Надстройки* и щелкните по кнопке **Состав...** – откроется окно **Microsoft Office 97-Надстройки** (рис. 1.4);

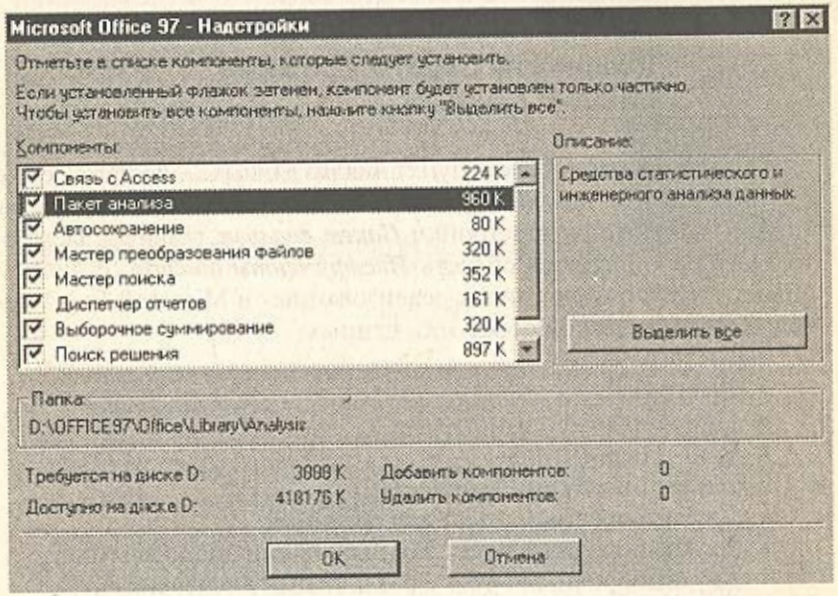

Рис. 1.4

4) в списке окна **Microsoft Office 97-Надстройки** найдите элемент *Пакет анализа*, поставьте рядом с ним «галку» и щелкните по кнопке **OK**;

5) в каждом «родительском» окне щелкайте по кнопке **OK**, в окне **Microsoft Office 97-Сопровождение** – по кнопке **Далее**, отвечайте затем по умолчанию на сообщения, выводимые программой установки.

Примечания: 1. Рекомендуем при установке надстройки ***Пакет анализа*** установить сразу и все другие надстройки Microsoft Excel. Они значительно расширят возможности программы, а при этом займут на винчестере совсем немного места – около 4 Мбайт.

2. После окончания установки в папке D:\Office97\Office\ Library\ \Analysis* появится файл надстройки analysis32.xll.

Итак, мы рассмотрели все возможные ситуации, связанные с установкой надстройки ***Пакет анализа***, и переходим к знакомству с технологией работы в режиме «Анализ данных».

1.1.2. Технология работы в режиме «Анализ данных»

Выберем в меню ***Сервис*** пункт **Анализ данных...**, появится окно с одноименным названием (рис. 1.5). Это окно – по существу «центр управления» надстройки ***Пакет анализа***, главным элементом которого является область ***Инструменты анализа***. В данной области представлен список реализованных в Microsoft Excel методов статистической обработки данных:

- «Гистограмма»;
- «Выборка»;
- «Описательная статистика»;
- «Ранг и персентиль»;
- «Генерация случайных чисел»;
- «Двухвыборочный *z*-тест для средних»;
- «Двухвыборочный *t*-тест с одинаковыми дисперсиями»;
- «Двухвыборочный *t*-тест с различными дисперсиями»;
- «Двухвыборочный *F*-тест для дисперсий»;
- «Парный двухвыборочный *t*-тест для средних»;
- «Однофакторный дисперсионный анализ»;
- «Двухфакторный дисперсионный анализ без повторений»;
- «Двухфакторный дисперсионный анализ с повторениями»;

* Логический диск и главная родительская папка (в нашем примере D:\Office 97), в которой располагается пакет Microsoft Office, могут иметь другие имена. Они задаются при первой инсталляции пакета на компьютер.

- «Ковариация»;
- «Корреляция»;
- «Регрессия»;
- «Скользящее среднее»;
- «Экспоненциальное сглаживание»;
- «Анализ Фурье».

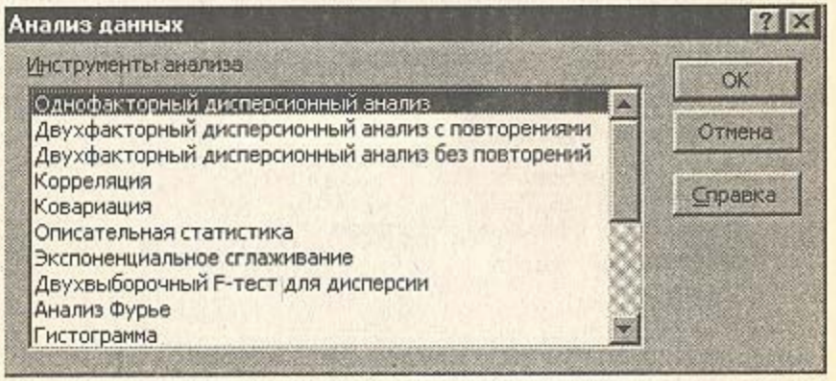

Рис. 1.5

Каждый из перечисленных методов реализован в виде отдельного режима работы, для активизации которого необходимо выделить соответствующий метод указателем мыши и щелкнуть по кнопке **OK**. После появления диалогового окна вызванного режима можно приступать к работе.

Диалоговое окно каждого режима включает в себя элементы управления (поля ввода, раскрывающиеся списки, флажки, переключатели и т. п.), которые задают определенные параметры выполнения режима (в качестве примера на рис. 1.6 изображено диалоговое окно режима «Гистограмма»).

Одна часть параметров является специфической и присуща только одному (или малой группе) режиму работы. Назначение таких параметров будет рассмотрено при изучении технологий работы с соответствующими режимами.

Другая часть параметров универсальна и присуща всем (или подавляющему большинству) режимам работы. Элементами управления, задающими такие параметры, являются:

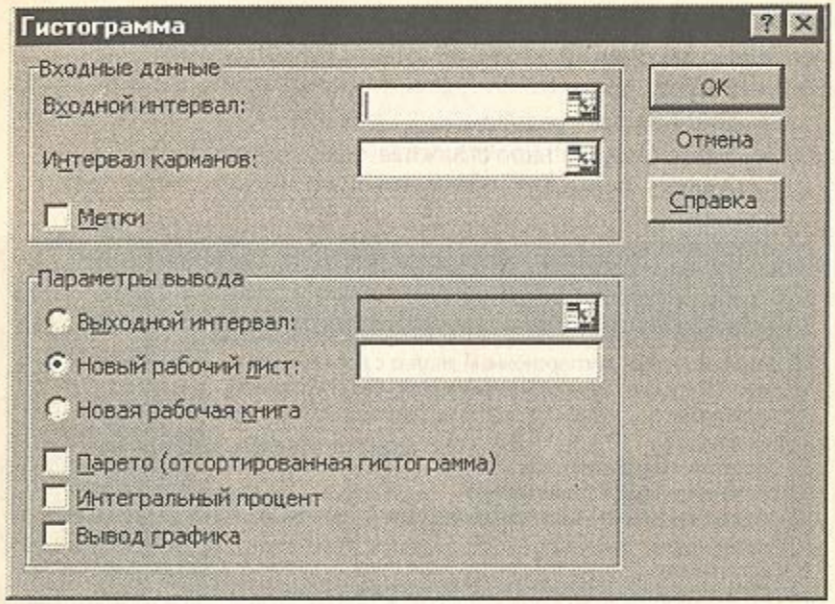

Рис. 1.6

1. Поле *Входной интервал* – вводится ссылка на ячейки, содержащие анализируемые данные.

2. Переключатель *Группирование* – устанавливается в положение *По столбцам* или *По строкам* в зависимости от расположения данных во входном диапазоне.

3. Флажок *Метки* – устанавливается в активное состояние, если первая строка (столбец) во входном диапазоне содержит заголовки. Если заголовки отсутствуют, флажок следует деактивизировать. В этом случае будут автоматически созданы стандартные названия для данных выходного диапазона.

4. Переключатель *Выходной интервал/Новый рабочий лист/Новая рабочая книга*.

В положении *Выходной интервал* активизируется поле, в которое необходимо ввести ссылку на левую верхнюю ячейку выходного диапазона. Размер выходного диапазона будет определен автоматически, и на экране появится сообщение в случае возможного наложения выходного диапазона на исходные данные.

В положении *Новый рабочий лист* открывается новый лист, в который начиная с ячейки A1 вставляются результаты анализа. Если необходимо задать имя открываемого нового рабочего листа, введите его имя в поле, расположенное напротив соответствующего положения переключателя.

В положении *Новая рабочая книга* открывается новая книга, на первом листе которой начиная с ячейки A1 вставляются результаты анализа.

Особенности технологии работы в каждом режиме подробно описаны в следующих главах.

1.2. Первое знакомство со статистическими функциями MS Excel

1.2.1. Работа с мастером функций

Наряду с надстройкой ***Пакет анализа*** в практике статистической обработки могут широко применяться статистические функции Microsoft Excel. В состав Excel входит библиотека, содержащая 78 статистических функций, ориентированных на решение самых различных задач прикладного статистического анализа. Причем одну часть статистических функций можно рассматривать как своего рода элементарные составляющие того или иного режима надстройки ***Пакет анализа***, другую часть – как уникальные функции, не дублирующиеся в надстройке ***Пакет анализа***. Тем не менее функции, входящие и в первую часть, и во вторую часть, имеют самостоятельное значение и могут применяться автономно при решении конкретных статистических задач.

Работать со статистическими функциями Excel, как, впрочем, и с функциями из других категорий, удобнее всего с помощью мастера функций.

При работе с мастером функций необходимо сначала выбрать саму функцию, а затем задать ее отдельные аргументы. Запустить мастер функций можно командой **Функция...** из меню ***Вставка,*** или щелчком по кнопке вызова мастера функций (рис. 1.7), или активизацией комбинации клавиш **Shift+F3**.

Рис. 1.7

Для упрощения работы с мастером отдельные функции сгруппированы по тематическому признаку. Тематические категории представлены в области ***Категория*** (рис. 1.8). В категории ***Полный алфавитный перечень*** содержится список всех доступных в программе функций. К категории ***10 недавно использовавшихся*** относятся десять применявшихся последними функций. Поскольку пользователь во время работы применяет ограниченное число функций, то с помощью этой категории можно получить быстрый доступ к тем из них, которые необходимы в повседневной работе.

Чтобы задать статистическую функцию, сначала необходимо выбрать категорию ***Статистические***. При перемещении строки выделения по списку функций под областями ***Категория*** и ***Функция*** будет представлен пример, иллюстрирующий способ задания выбранной статистической функции с краткой информацией о ней.

Если краткой информации недостаточно, щелкните в диалоговом окне по кнопке **Справка** (или воспользуйтесь клавишей **F1**). На экране появится помощник и предложит помощь. Щелкните по кнопке **Справка по выделенной функции**, и на экране будет представлена соответствующая страница справочной подсистемы.

После выбора функции щелкните по кнопке **ОК** для перехода в следующее диалоговое окно мастера функций, в котором должны быть заданы аргументы. В этом диалоговом окне мастер подсказывает пользователю, какие аргументы следует указать обязательно (обязательные аргументы), а какие – опционально (необязательные аргументы).

Задать аргументы можно различными способами, наиболее удобные из них предлагает помощник. После задания всех аргу-

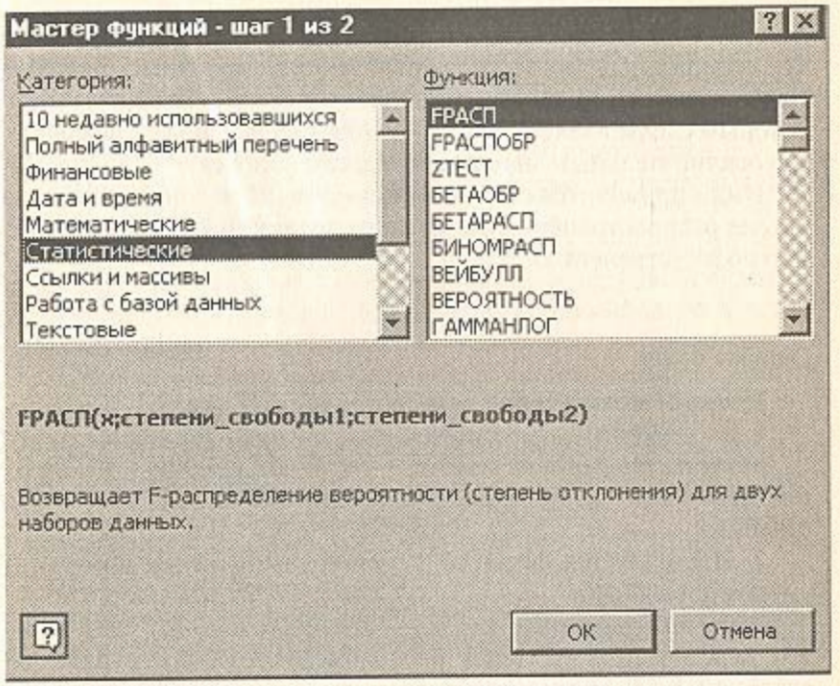

Рис. 1.8

ментов функции щелкните по кнопке **ОК**, чтобы в ячейке появились результаты выполнения функции.

Более подробно о каждой статистической функции можно узнать из следующих глав.

1.2.2. Виды ошибок при задании формул

Формула в Microsoft Excel представляет собой синтаксическую конструкцию, начинающуюся со знака равенства (=) и предназначенную для обработки данных с последующим помещением результатов обработки в ячейку, где записана сама формула. Формула может содержать одну или несколько функций, связанных между собой арифметическими операторами или вложенных друг в друга. Если при задании формулы были допущены ошибки, результатом ее вычисления будет так называемое значение ошибки,

которое появится в ячейке. В зависимости от вида ошибки в ячейке, содержащей формулу, записываются различные значения. Первым символом значения ошибки является символ диез (#), за которым следует текст. Текст значения ошибки может завершаться восклицательным знаком или знаком вопроса.

Ниже приводится список значений ошибок с пояснением наиболее распространенных причин их возникновения и указанием мер по их устранению.

Ошибка #####

Причины возникновения ошибки

1. Вводимое числовое значение не умещается в ячейке.

Меры по устранению ошибки – увеличьте ширину столбца путем перемещения границы, расположенной между заголовками столбцов.

2. Используется формула, результат выполнения которой не умещается в ячейке.

Меры по устранению ошибки – увеличьте ширину столбца путем перемещения границы, расположенной между заголовками столбцов. Кроме того, можно изменить формат числа ячейки, для чего следует выбрать команду **Ячейки...** в меню ***Формат***, затем – вкладку ***Число*** и указать другой формат.

3. При определении числа дней между двумя датами, а также количества часов между двумя временными промежутками получается отрицательное значение.

Меры по устранению ошибки – введите правильно формулу, чтобы число дней (или часов) было положительным числом.

Ошибка #ЗНАЧ!

Причины возникновения ошибки

1. Вместо числового или логического значения введен текст, и Microsoft Excel не может преобразовать его к нужному типу данных.

Меры по устранению ошибки – проверьте в формуле правильность задания типов аргументов. Например, если в ячейке A1 содержится число 5, в ячейке B1 – текстовое значение «Привет», а в

ячейке C1 – формула =A1+B1, то в ячейке C1 будет выведена ошибка #ЗНАЧ!. Если все же необходимо сложить два таких значения, то следует использовать функцию СУММ (функция СУММ игнорирует текстовые значения). Для рассматриваемой ситуации функция = СУММ(A1:B1) рассчитывает значение 5.

2. После ввода или редактирования формулы массива нажимается клавиша **Enter.**

Меры по устранению ошибки – для редактирования формулы укажите ячейку или диапазон ячеек, содержащих формулу массива, нажмите клавишу **F2**, а затем – комбинацию клавиш **Ctrl+Shift +Enter.**

3. Использована неправильная размерность матрицы данных в одной из матричных функций листа.

Меры по устранению ошибки – укажите правильную размерность матрицы данных.

Ошибка #ДЕЛ/0!

Причины возникновения ошибки

1. В качестве делителя используется ссылка на ячейку, содержащую нулевое или пустое значение (если аргумент является пустой ячейкой, то ее содержимое интерпретируется как нуль). Такая ситуация чаще всего возникает случайно: например, если ячейка содержит формулу =A1/B1, а содержимое ячейки B1 по какой-либо причине было удалено.

Меры по устранению ошибки – измените ссылку или введите ненулевое значение в ячейку, используемую в качестве делителя. Кроме того, в качестве делителя можно ввести значение #Н/Д. В этом случае ошибка #ДЕЛ/0! сменится на #Н/Д, указывающую, что значение делителя не определено.

2. В формуле содержится явное деление на нуль, например = 5/0.

Меры по устранению ошибки – исправьте формулу.

Ошибка #ИМЯ?

Причины возникновения ошибки

1. Используемое в формуле имя было удалено или не было определено.

Меры по устранению ошибки – определите имя. Для этого выберите команду **Имя** в меню ***Вставка***, а затем – команду

Создать.... Кроме того, команда **Создать...** используется для добавления имени, отсутствующего в списке.

2. Имеется ошибка в написании имени.

Меры по устранению ошибки – исправьте написание имени. Чтобы вставить правильное имя в формулу, выделите имя в строке формул, выберите команду **Имя** в меню ***Вставка***, а затем – команду **Вставить**. На экране появится диалоговое окно **Вставка имени**. Выделите нужное имя и щелкните по кнопке **ОК**.

3. Имеется ошибка в написании имени функции.

Меры по устранению ошибки – исправьте написание имени функции вручную или вставьте функцию с помощью мастера функций.

4. В формулу введен текст, не заключенный в двойные кавычки. Microsoft Excel пытается распознать такой текст как имя, хотя это не предполагается.

Меры по устранению ошибки – заключите текст формулы в двойные кавычки. Например, если в ячейке A1 содержится значение 200, а в ячейке B1 – формула =«Итого:»&A1, то в ячейке B1 будет выведен результат Итого:200.

5. В ссылке на диапазон ячеек пропущен знак двоеточия (:).

Меры по устранению ошибки – исправьте формулу так, чтобы во всех ссылках на диапазон ячеек использовался знак двоеточия (:), например =СУММ(A1:C10).

Ошибка #Н/Д

Значение ошибки #Н/Д (Неопределенные данные) помогает предотвратить использование ссылки на пустую ячейку. Введите в ячейки листа значение #Н/Д, если они должны содержать данные, но в настоящий момент эти данные отсутствуют. Формулы, ссылающиеся на эти ячейки, тоже будут иметь значение #Н/Д.

Причины возникновения ошибки

1. Для функций ГПР, ПРОСМОТР, ПОИСКПОЗ или ВПР (функции ссылки и автоподстановки) задан недопустимый аргумент *искомое__значение*.

Меры по устранению ошибки – задайте правильный аргумент *искомое__значение*, например значение или ссылку, но не диапазон ссылок.

2. Функции ВПР или ГПР используются для обработки неотсортированной таблицы.

Меры по устранению ошибки – по умолчанию для функций просмотра таблиц сведения должны располагаться в возрастающем порядке (аргумент *интервальный__просмотр* опущен или имеет значение ИСТИНА). Чтобы найти искомое значение в неотсортированной таблице, установите для аргумента *интервальный__просмотр* значение ЛОЖЬ.

3. В формуле массива используется аргумент, не соответствующий размеру диапазона, определяющегося числом строк и столбцов.

Меры по устранению ошибки – если формула массива введена в несколько ячеек, проверьте диапазон ссылок формулы на соответствие числу строк и столбцов или введите формулу массива в недостающие ячейки. Например, если формула массива введена в первые 15 ячеек столбца C (C1:C15), а сама формула ссылается на первые 10 ячеек столбца A (A1:A10), то в ячейках C11:C15 будет отображаться ошибка #Н/Д. Чтобы исправить эту ошибку, уменьшите диапазон в формуле (например, C1:C10) или увеличьте диапазон, на который ссылается формула (например, A1:A15).

4. Не заданы один или несколько аргументов стандартной или пользовательской функции листа.

Меры по устранению ошибки – задайте все необходимые аргументы функции.

5. Используется пользовательская функция, обращение к которой приводит к ошибке.

Меры по устранению ошибки – проверьте, что книга, использующая функцию листа, открыта, и убедитесь в правильности работы функции (проведите отладку в редакторе VBA).

Ошибка #ССЫЛКА!

Причина возникновения ошибки

Ячейки, на которые ссылаются формулы, были удалены или в эти ячейки было помещено содержимое других скопированных ячеек.

Меры по устранению ошибки – измените формулы или сразу же после удаления или вставки скопированного восстановите прежнее содержимое ячеек с помощью кнопки **Отменить**.

Ошибка #ЧИСЛО!

Причины возникновения ошибки

1. В функции с числовым аргументом используется неприемлемый аргумент.

Меры по устранению ошибки – проверьте правильность использования в функции аргументов.

2. Задана функция (например, статистическая функция СТЬЮДРАСПОБР), при вычислении которой используется итерационный процесс. При этом итерационный процесс не сходится и результат не может быть получен.

Меры по устранению ошибки – используйте другое начальное приближение для этой функции.

3. Введена формула, рассчитывающая числовое значение, которое слишком велико или слишком мало, чтобы его можно было представить в Microsoft Excel.

Меры по устранению ошибки – измените формулу так, чтобы в результате ее вычисления получалось число, попадающее в диапазон от $-1 \cdot 10^{307}$ до $1 \cdot 10^{307}$. Например, число 200 является слишком большим, чтобы быть использованным в качестве аргумента функции ФАКТР (функция вычисления факториала числа), поэтому формула =ФАКТР(200) помещает в ячейку значение ошибки #ЧИСЛО!.

Ошибка #ПУСТО!

Причина возникновения ошибки

Использован оператор, задающий пересечение диапазонов, не имеющих общих ячеек.

Меры по устранению ошибки – задайте правильно размерность пересекающихся диапазонов или не используйте оператор пересечения, если диапазоны не являются таковыми.

В Microsoft Excel оператором пересечения диапазонов является пробел ( ). Например, диапазоны A1:A5 и B1:B5 содержат массивы единиц. В этом случае формула =СУММ(A1:A5; B1:B5) будет выдавать значение ошибки #ПУСТО!, а формула =СУММ(A1:A5; A1:B3) рассчитает значение 3. Для суммирования непересекающихся диапазонов A1:A5 и B1:B5 необходимо воспользоваться стандартной синтаксической конструкцией функции СУММ, т. е. =СУММ(A1:A5;B1:B5), которая рассчитает значение 10.

ГЛАВА 2
Гистограмма

2.1. Краткие сведения из теории статистики

Результаты сводки и группировки материалов статистического наблюдения оформляются в виде таблиц и статистических рядов распределения.

Статистический ряд распределения представляет собой упорядоченное распределение единиц изучаемой совокупности по определенному варьирующему признаку. Он характеризует состояние (структуру) исследуемого явления, позволяет судить об однородности совокупности, границах ее изменения, закономерностях развития наблюдаемого объекта. Построение рядов распределения является составной частью сводной обработки статистической информации.

В зависимости от признака, положенного в основу образования ряда распределения, различают *атрибутивные* и *вариационные* ряды распределения [8, 12]. Последние, в свою очередь, в зависимости от характера вариации признака делятся на *дискретные* (*прерывные*) и *интервальные (непрерывные)* ряды распределения.

Удобнее всего ряды распределения анализировать с помощью их графического изображения, позволяющего судить о форме распределения. Наглядное представление о характере изменения частот вариационного ряда дают полигон и гистограмма.

Полигон используется для изображения *дискретных* вариационных рядов. При построении полигона в прямоугольной системе координат по оси абсцисс в одинаковом масштабе откладываются ранжированные значения варьирующего признака, а по оси ординат наносится шкала частот, т. е. число случаев, в которых встретилось то или иное значение признака*. Полученные на пе-

* На оси ординат могут наноситься не только значения частот, но и частостей вариационного ряда. Частостями называют частоты, выраженные в долях единицы или в процентах к итогу. Соответственно сумма частостей равна 1 или 100%. В математической статистике наряду с термином «частость» также широко употребляется термин «статистическая вероятность».

ресечении абсцисс и ординат точки соединяют прямыми линиями, в результате чего получают ломаную линию, называемую полигоном частот.

Например, в табл. 2.1 представлено распределение жилого фонда городского района по типу квартир [12]. Построим полигон для данного распределения.

Таблица 2.1

	C	D	E
52	**Распределение жилого фонда городского района по типу квартир**		
53	№ п/п	Группы квартир по числу комнат	Число квартир, тыс. ед.
54	1	1	10
55	2	2	35
56	3	3	30
57	4	4	15
58	5	5	5
59	ВСЕГО		95

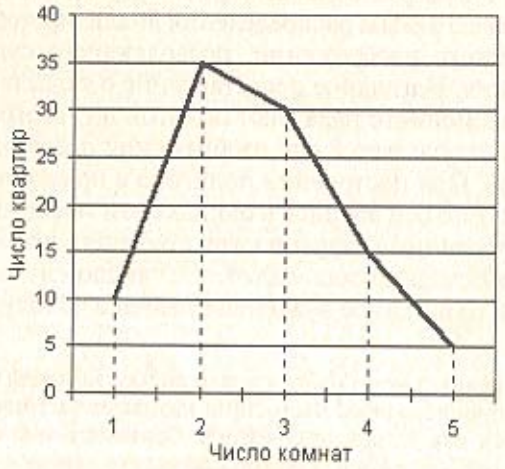

Рис. 2.1

Для построения полигона воспользуемся мастером диаграмм Microsoft Excel (режим «График») и получим полигон (рис. 2.1).

Для изображения *интервальных* вариационных рядов распределений применяются *гистограммы*. При этом на оси абсцисс откладываются значения интервалов, а частоты изображаются прямоугольниками, построенными на соответствующих интервалах. В результате получается гистограмма – график, на котором ряд распределения представлен в виде смежных друг с другом областей.

На рис. 2.2 показана построенная с помощью мастера диаграмм гистограмма интервального ряда распределения, приведенного в табл. 2.2 [12].

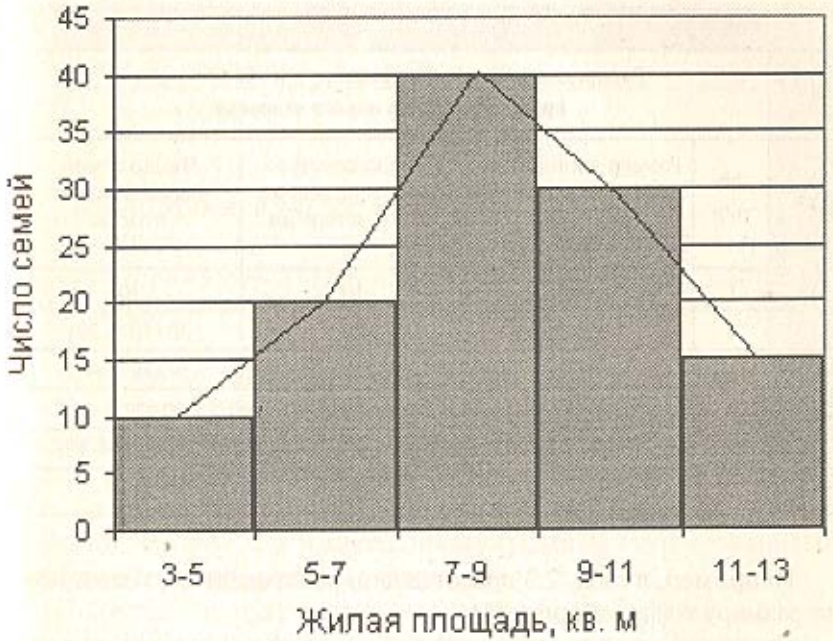

Рис. 2.2

При необходимости гистограмма интервального ряда распределения может быть преобразована в полигон. Для этого нужно середины верхних сторон прямоугольников соединить прямыми линиями (ломаная линия на рис. 2.2).

В рассмотренном распределении (*см.* табл. 2.2) интервалы имеют одинаковую величину, поэтому высота столбиков гистограммы пропорциональна частотам ряда распределения. При неравных интервалах это условие не соблюдается, что не позволяет правильно оценить характер распределения по данному признаку. В подобных случаях для обеспечения необходимой сравнимости исчисляют *плотность статистического распределения*, т. е. определяют, сколько единиц в каждой группе приходится на единицу величины интервала.

Таблица 2.2

	B	C	D	E
12	**Распределение семей по размеру жилой площади, приходящейся на одного человека**			
13	№ п/п	Размер жилой площади, приходящейся на одного человека, $м^2$	Число семей с данным размером жилой площади	Число семей нарастающим итогом
14	1	3–5	10	10
15	2	5–7	20	30 (10 + 20)
16	3	7–9	40	70 (30 + 40)
17	4	9–11	30	100 (70 + 30)
18	5	11–13	15	115 (100 + 15)
19	ВСЕГО		115	

Например, в табл. 2.3 представлено распределение магазинов по размеру товарооборота [8].

Сравнение частот отдельных групп показывает, что чаще всего встречаются магазины с товарооборотом 250–450 тыс. руб., что не является совсем верным. Для точной характеристики магазинов по товарообороту рассчитаем плотность распределения путем деления значений частот на величину интервала ({=D11:D15/ E11: E15}). Оказывается, что чаще всего встречаются магазины с товарооборотом 50–120 тыс. руб.

Таблица 2.3

	B	C	D	E	F
9	**Распределение магазинов по размеру товарооборота**				
10	№ п/п	Группы магазинов по размеру товарооборота, тыс. руб.	Число магазинов	Величина интервала, тыс. руб.	Плотность распределения (D/E)
11	1	До 50	25	50	0,5
12	2	50–120	45	70	0,64
13	3	120–250	65	130	0,5
14	4	250–450	80	200	0,4
15	5	450–980	20	530	0,04
16	ИТОГО		235		

При построении гистограммы вариационного ряда с неравными интервалами высоту прямоугольников определяют пропорционально не частотам, а показателям плотности распределения значений изучаемого признака в соответствующих интервалах.

В практике экономической работы нередко возникает потребность в преобразовании рядов распределения в *кумулятивные ряды*, строящиеся по *накопленным* частотам. С их помощью можно определять структурные средние (*см.* главу 3) и наблюдать за процессом концентрации изучаемого явления (*кривые Лоренца)*. На рис. 2.3 изображена кумулята для интервального ряда распределения, приведенного в табл. 2.2.

Полигон и кумулята дают начальное представление о *функции* и *плотности распределения случайной величины*. При этом полигон можно рассматривать в качестве статистического аналога плотности распределения, а кумуляту – в качестве статистического аналога функции распределения. Более подробно о функции и плотности распределения случайной величины, а также о задаваемых с их помощью теоретических распределениях *см.* в главе 6.

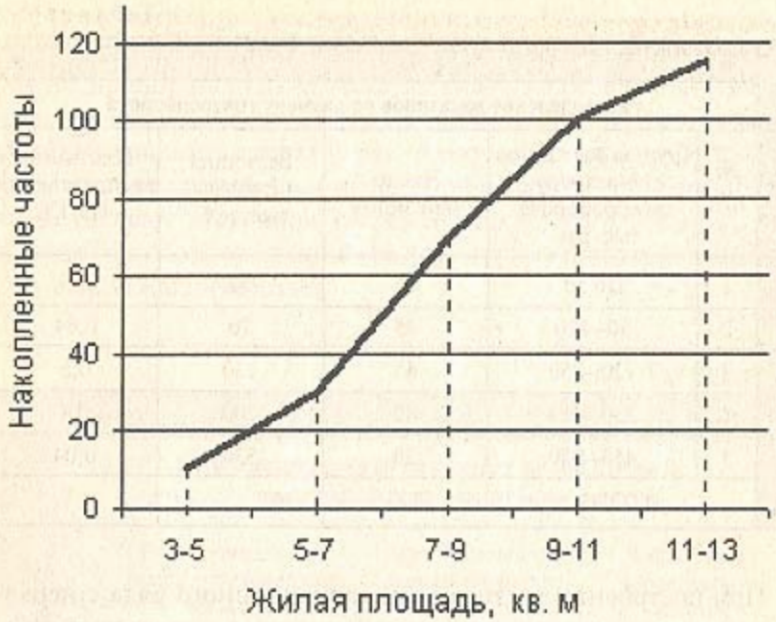

Рис. 2.3

2.2. Справочная информация по технологии работы

Режим «Гистограмма» служит для вычисления частот попадания данных в указанные границы интервалов, а также для построения гистограммы *интервального* вариационного ряда распределения.

В диалоговом окне данного режима (рис. 2.4) задаются следующие параметры:

1. *Входной интервал – см.* подразд. 1.1.2.

2. *Интервал карманов* (необязательный параметр) – вводится ссылка на ячейки, содержащие набор граничных значений, определяющих интервалы (карманы). Эти значения должны быть введены в возрастающем порядке. В Microsoft Excel вычисляется число попаданий данных в сформированные интервалы, причем границы интервалов являются строгими нижними границами и нестрогими верхними: $a < x \le b$.

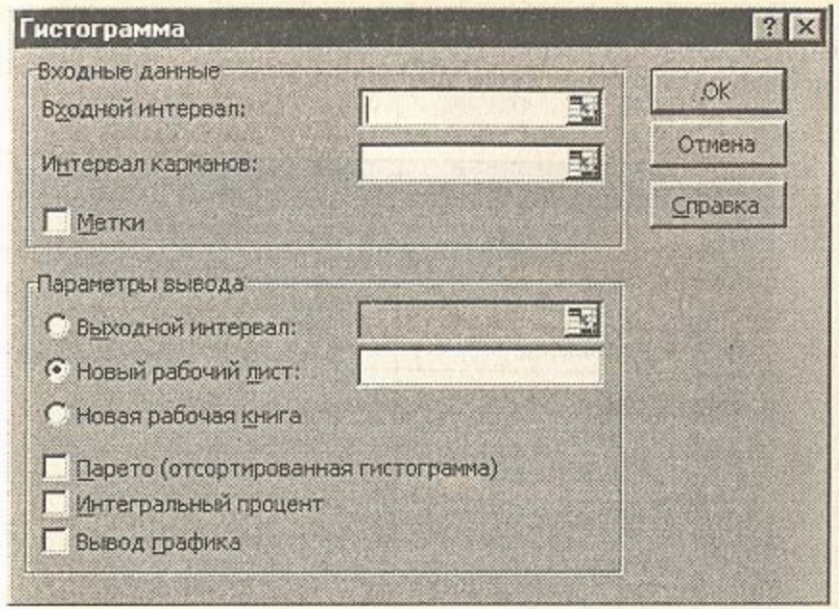

Рис. 2.4

Если диапазон карманов не был введен, то набор интервалов, равномерно распределенных между минимальным и максимальным значениями данных, будет создан автоматически.

3. *Метки – см.* подразд. 1.1.2.

4. *Выходной интервал/Новый рабочий лист/Новая рабочая книга – см.* подразд. 1.1.2.

5. *Парето* (*отсортированная гистограмма*) – устанавливается в активное состояние, чтобы представить данные в порядке убывания частоты. Если флажок снят, то данные в выходном диапазоне будут приведены в порядке следования интервалов.

6. *Интегральный процент* – устанавливается в активное состояние для расчета выраженных в процентах *накопленных* частот (*накопленных частостей)* и включения в гистограмму графика кумуляты.

7. *Вывод графика* – устанавливается в активное состояние для автоматического создания встроенной диаграммы на листе, содержащем выходной диапазон.

Пример 2.1. Общий объем розничного товарооборота по районам Ярославской области за 1998 г. приведен в табл. 2.4 [2], сформированной на рабочем листе Microsoft Excel.

Таблица 2.4

	B	C
38	**Объем розничного товарооборота по районам Ярославской области за 1998 г.**	
39	Район	Товарооборот, млн руб.
40	Большесельский	31,0
41	Борисоглебский	38,5
42	Брейтовский	34,0
43	Гаврилов-Ямский	87,6
44	Даниловский	139,6
45	Любимский	46,0
46	Мышкинский	46,0
47	Некоузский	76,6
48	Некрасовский	68,3
49	Первомайский	41,1
50	Переславский	93,7
51	Пошехонский	80,9
52	Ростовский	52,6
53	Рыбинский	76,3
54	Тутаевский	45,8
55	Угличский	28,5
56	Ярославский	190,5

По набору данных (*см.* табл. 2.4) необходимо построить гистограмму и кумуляту.

Для решения задачи воспользуемся режимом работы «Гистограмма». Значения параметров, установленных в диалоговом окне **Гистограмма**, показаны на рис. 2.5. Частоты и накопленные частости, рассчитанные в данном режиме, представлены в табл. 2.5, а построенные гистограмма и кумулята изображены на рис. 2.6.

Поясним подробнее порядок расчета накопленных частостей (*см.* в табл. 2.5 графу «Интегральный %»). На основании частот

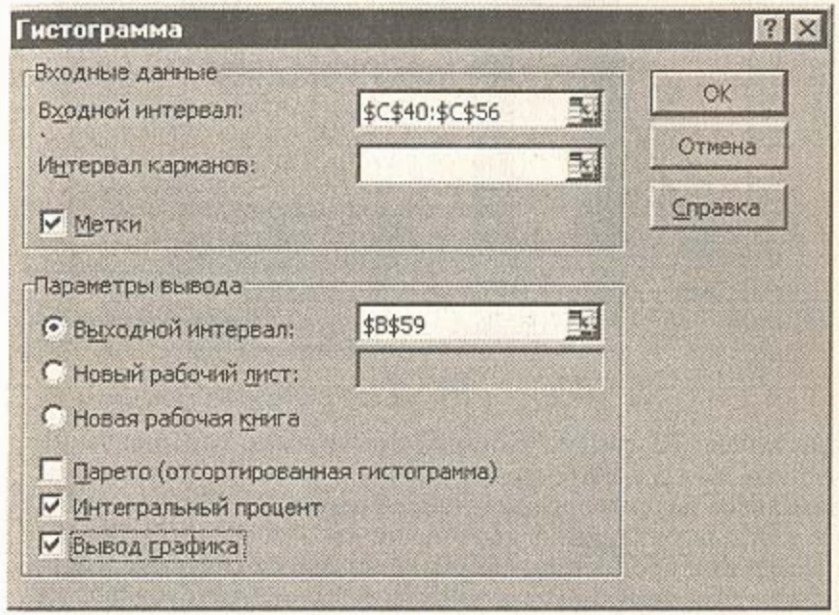

Рис. 2.5

Таблица 2.5

	B	C	D
59	*Карман*	*Частота*	*Интегральный %*
60	28,5	1	5,88%
61	69	9	58,82%
62	109,5	5	88,24%
63	150	1	94,12%
64	Еще	1	100,00%

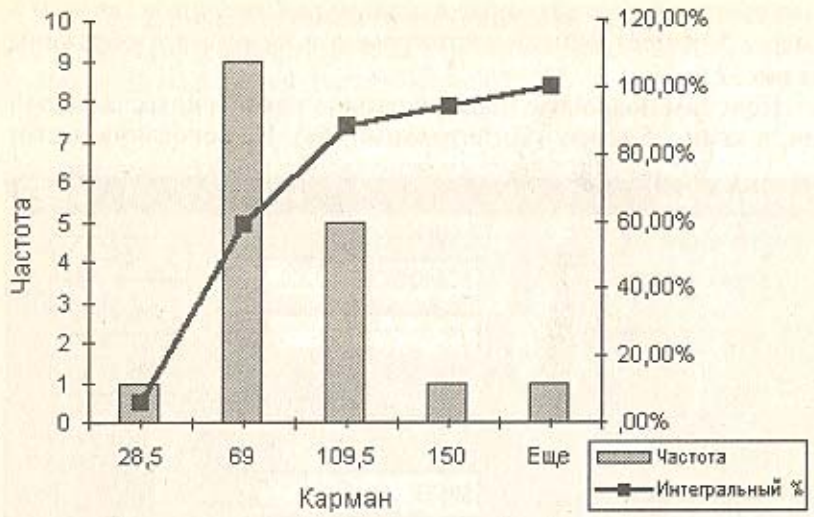

Рис. 2.6

(*см.* в табл. 2.5 графу «Частота») рассчитываются накопленные частоты. Каждое значение накопленной частоты делится на максимальное накопленное значение, в результате чего получаются частости, выраженные в долях единицы. После преобразования последних к процентному формату получаем окончательный результат. Промежуточные и заключительные итоги вычислений сведены в табл. 2.6.

Как правило, гистограммы изображаются в виде *смежных* прямоугольных областей, поэтому столбики гистограммы на рис. 2.6 целесообразно расширить до соприкосновения друг с другом. Для этого на панели инструментов ***Диаграмма*** необходи-

Таблица 2.6

Частота	Накопленная частота	Частость	Частость, %
1	1	0,0588	5,88
9	10	0,5882	58,82
5	15	0,8824	88,24
1	16	0,9412	94,12
1	17	1,0000	100,00

мо в раскрывающемся списке элементов диаграммы выбрать элемент *Ряд 'Частота'*, после чего щелкните по кнопке **Формат рядов данных**. В появившемся одноименном диалоговом окне необходимо активизировать вкладку ***Параметры*** и в поле *Ширина зазора* установить значение 0. После указанных преобразований гистограмма примет стандартный вид (рис. 2.7).

Внимание! В примере 2.1 величина интервала определялась автоматически в соответствии с формулой

$$h = \frac{x_{max} - x_{min}}{\{n\} - 1}, \qquad (2.1)$$

где h – величина равного интервала;
x_{max}, x_{min} – соответственно максимальное и минимальное значения признака в совокупности;
$\{n\}$ – округленное оптимальное число групп, определяемое по формуле Стерджесса $n = 1 + 3{,}322 \cdot \lg N$ (N – число единиц совокупности).

Так, для примера 2.1 имеем:

$$n = 1 + 3{,}322 \cdot \lg 17 \approx 5{,}09;$$

$$h = \frac{190{,}5 - 28{,}5}{5 - 1} = 40{,}5.$$

Примечание. Формула (2.1) используется только при работе в режиме «Гистограмма». В других случаях следует применять формулу

$$h = \frac{x_{max} - x_{min}}{n}.$$

В режиме работы «Гистограмма» пользователь может самостоятельно задать величину интервалов ряда (параметр *Интервал карманов* диалогового окна **Гистограмма**). В случае если заданные интервалы будут не равны между собой, то сгенерированная гистограмма будет представлять собой обычную столбиковую диаграмму, в которой частоты попадания в интервал не связаны с его размером, что не позволит правильно оценить характер распределения изучаемого явления. Во избежание подобных ошибок рекомендуется задавать интервалы одинаковой величины или пользоваться режимом автоматического формирования интервалов.

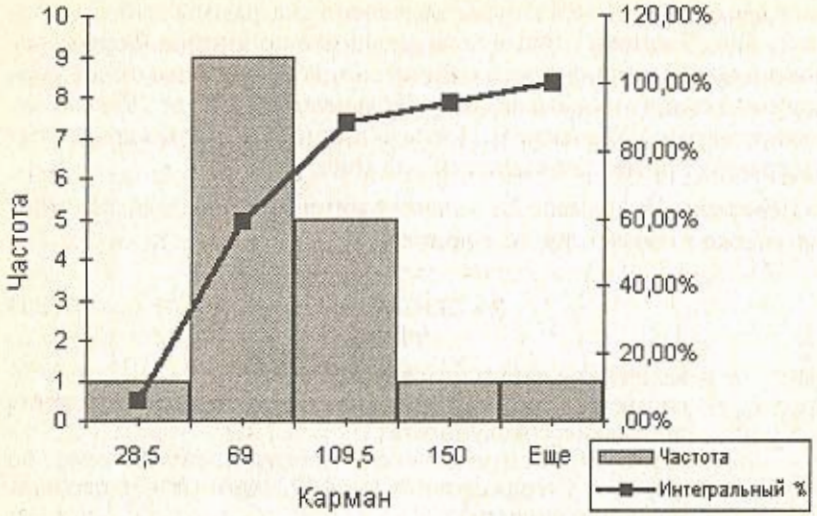

Рис. 2.7

2.3. Статистические функции, связанные с режимом «Гистограмма»

Функция ЧАСТОТА

См. также СЧЕТ, СЧЕТЗ.

Синтаксис:

ЧАСТОТА (массив __ данных; массив __ карманов).

Результат:

Вычисляет для множества исходных данных число значений, попадающих в заданные интервалы, т. е. частоты статистического распределения.

Аргументы:

• *массив __ данных:* массив множества данных, для которых вычисляются частоты. Если *массив __ данных* не содержит значений, то функция ЧАСТОТА помещает в ячейки массив нулей;

• *массив __ карманов:* массив интервалов, в которые группируются значения аргумента *массив __ данных*. Если *массив __ карманов* не содержит значений, то функция ЧАСТОТА рассчитывает количество элементов в аргументе *массив __ данных*.

Замечания:

• функция ЧАСТОТА вводится как формула массива после выделения интервала смежных ячеек, в которые нужно поместить рассчитываемый массив распределения;

• количество элементов в результирующем массиве на единицу больше количества элементов в аргументе *массив __ карманов*;

• функция ЧАСТОТА игнорирует пустые ячейки и тексты.

Математико-статистическая интерпретация:

Функция ЧАСТОТА рассчитывает для множества исходных данных массив частот, соответствующих числу появлений значений в заданных интервалах. Интервалы значений задаются в аргументе *массив __ карманов*, причем границы интервалов являются строгими нижними границами и нестрогими верхними: $a < x \le b$.

Примечание. Если требуется задать интервал с другим характером границ (например, нестрогими нижними границами и строгими верхними: $a \le x < b$), то в этом случае необходимо воспользоваться функцией СЧЕТЕСЛИ.

♦ В примере 2.1 значения частот (*см.* в табл. 2.5 графу *Частота*) рассчитываются по формуле массива

{=ЧАСТОТА(C40:C56;B60:B63)},

где диапазон C40:C56 содержит массив исходных данных (*см.* табл. 2.4) , а диапазон B60:B63 – массив автоматически рассчитываемых границ интервалов (*см.* в табл. 2.5 графу *Карман*).

ГЛАВА 3 Выборка

3.1. Краткие сведения из теории статистики

Методология исследования массовых статистических явлений в зависимости от полноты охвата изучаемого объекта (явления) различает *сплошное* и *несплошное* наблюдение [8, 12]. Разновидностью несплошного наблюдения является выборочное, которое в

условиях развития современных рыночных отношений находит все более широкое применение.

Под *выборочным* наблюдением понимается метод статистического исследования, при котором обобщающие показатели изучаемой совокупности устанавливаются по некоторой ее части на основе положений случайного отбора. При выборочном методе обследованию подвергается сравнительно небольшая часть всей изучаемой совокупности, получившая название *выборочной совокупности* или просто *выборки*.

Выборка должна быть *представительной (репрезентативной)*, чтобы по ней можно было судить о генеральной совокупности. Репрезентативность означает, что объекты выборки достаточно хорошо представляют генеральную совокупность. Заметим, что при отборе объектов могут сыграть роль личные мотивы или психологические факторы, о которых исследователь, проводящий выборку, и не подозревает. При этом выборка, как правило, не будет репрезентативной.

Предупреждение систематических (тенденциозных) ошибок выборочного обследования достигается в результате применения научно обоснованных способов формирования выборочной совокупности, в зависимости от которых выборка может быть [12]:

- собственно-случайной;
- механической;
- типической;
- серийной;
- комбинированной.

В табличном процессоре Microsoft Excel реализована собственно-случайная выборка.

Собственно-случайная выборка состоит в том, что выборочная совокупность образуется в результате случайного (непреднамеренного) отбора отдельных единиц из генеральной совокупности. Именно принцип случайности попадания любой единицы генеральной совокупности в выборку предупреждает возникновение систематических (тенденциозных) ошибок выборки.

Собственно-случайная выборка может быть осуществлена по схемам *повторного* и *бесповторного* отбора. Повторный отбор предполагает возможность включения в выборку одного и того же элемента генеральной совокупности два раза и более. Бесповторный отбор исключает такую возможность. В Microsoft Excel реализована схема *повторного отбора*.

На практике, особенно при большом объеме генеральной совокупности, для организации собственно-случайной выборки часто используют таблицу случайных чисел или генератор случайных чисел (*см.* подробнее в главе 6). В Microsoft Excel выборка формируется на основе *генератора случайных чисел*.

Предположим, например, что для проверки качества изготовленных за месяц приборов требуется сформировать контрольную выборку из 10 изделий. Прибор имеет заводской номер, присваиваемый по порядку. Допустим, что было изготовлено 500 приборов с номерами от 7001 до 7500 включительно. Тогда для формирования случайной выборки необходимо сгенерировать 10 случайных чисел из диапазона 7001–7500. Такая выборка является случайной выборкой с повторением, так как некоторые номера могут повторяться, следовательно, приборы с этими номерами должны обследоваться дважды. Если же необходимо организовать случайную выборку без повторения, то вновь встретившееся число следует пропустить и сгенерировать его повторно.

Выборочный метод, обладая несомненным достоинством, состоящим в возможности значительно сократить время на контроль и получение основных статистических характеристик, приводит к появлению ошибки и уменьшению гарантии получения истинных характеристик генеральной совокупности. Данное обстоятельство особенно важно учитывать при формировании так называемых *малых выборок*. При этом достаточно сложной проблемой является определение необходимого (оптимального) объема выборки. В математической статистике доказывается, что необходимая численность *собственно-случайной повторной* выборки определяется выражением

$$n = \frac{t^2 \sigma^2}{\Delta_x^2},$$

где Δ_x – предельная ошибка выборки;

σ^2 – дисперсия генеральной совокупности;

t – коэффициент доверия (определяется в зависимости от того, с какой доверительной вероятностью надо гарантировать результаты выборочного обследования)*.

*Более подробно об этих статистических показателях *см.* в подразд. 4.2, 4.4.3, 6.3.1, 6.3.8.

Затруднительным моментом применения приведенной формулы на практике является расчет генеральной дисперсии σ^2. Для ее оценки пользуются или материалами предыдущих исследований, или производственно-техническими нормативами, или, если предыдущие варианты неосуществимы, проводят пробное обследование. По результатам пробного обследования оценивают значение генеральной дисперсии для последующего обоснования необходимого объема выборки.

3.2. Справочная информация по технологии работы

Режим «Выборка» служит для формирования выборки из генеральной совокупности на основе схемы *повторного собственно-случайного* отбора, а также из *периодических* данных. Генеральная совокупность рассматривается при этом в качестве входного диапазона.

В диалоговом окне данного режима (рис. 3.1) задаются следующие параметры:

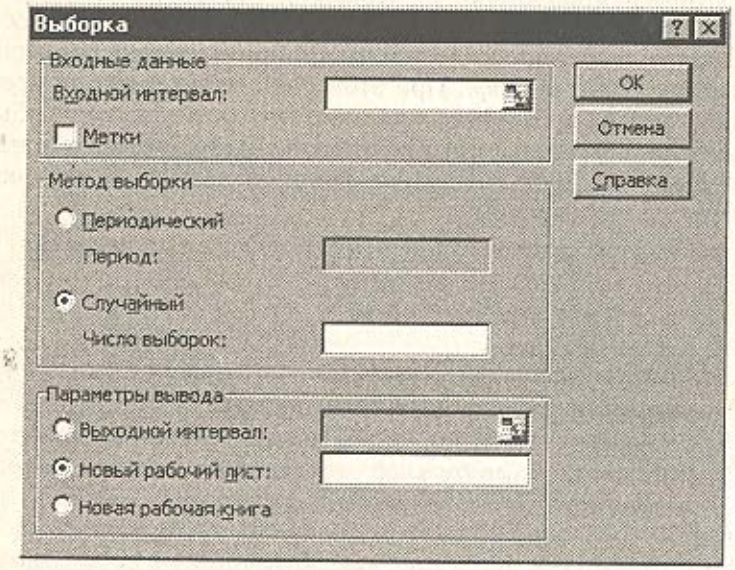

Рис. 3.1

1. *Входной интервал – см.* подразд. 1.1.2.
2. *Метки – см.* подразд. 1.1.2.
3. *Периодический/Случайный метод выборки.*

В положении *Периодический* активизируется поле *Период*, в которое необходимо ввести размер периодического интервала, в соответствии с которым будет сформирована выборка. Значение из генеральной совокупности, номер которого совпадает с номером, заданным в поле *Период*, и каждое последующее с номером, кратным периоду, будет скопировано в выходной столбец. Процесс создания выборки прекратится при достижении конца входного диапазона.

В положении *Случайный* активизируется поле *Число выборок*, в которое необходимо ввести число размещаемых в выходном столбце случайных значений. Позиция каждой извлекаемой переменной во входном диапазоне выбирается случайно, и любое исходное значение может быть выбрано более одного раза.

4. *Выходной интервал/Новый рабочий лист/Новая рабочая книга – см.* подразд. 1.1.2.

Пример 3.1. Фирма, торгующая бытовой техникой, решила для посетителей своего Web-сайта организовать лотерею по рассылке каталогов новой продукции. Для этого на сайте фирмы реализован счетчик посещений и предлагается (по желанию пользователя) заполнить электронный бланк с указанием своего почтового адреса. Отбор посетителей производится на основе показаний счетчика посещений за неделю. Для этого случайным образом отбираются пять показаний счетчика и проверяются соответствующие им регистрации посетителей. Если посетитель не указал своего адреса – каталог не высылается, в противном случае – высылается. При этом если одно и то же показание счетчика попало в выигрышную выборку несколько раз или несколько «выигрышных визитов» на сайт осуществил один и тот же посетитель, каталог высылается по одному и тому же адресу в соответствующем количестве экземпляров.

Рассмотрим следующую ситуацию. За последнюю неделю на сайте фирмы было зарегистрировано 25 посещений (показания счетчика увеличились с 360 до 385), информация по которым приведена в табл. 3.1, сформированной на рабочем листе Microsoft Excel.

Таблица 3.1

	B	C
2	Номер посещения	Информация о регистрации адреса
3	361	Адрес не указан
4	362	100050, г. Москва, Воздвиженка 17, 43
5	363	120005, г. Санкт-Петербург, Детская 12, 26
6	364	672007, г. Чита, Бунина 123, 7
7	365	250038, г. Тамбов, Державина 6, 75
8	366	Адрес не указан
9	367	Адрес не указан
10	368	340060, г. Саратов, Некрасова 46, 90
11	369	Адрес не указан
12	370	100050, г. Москва, Молодогвардейская 57, 12
13	371	100075, г. Москва, Варшавское шоссе 157, 20
14	372	460020, г. Новосибирск, академика Харитона 67, 34
15	373	Адрес не указан
16	374	325076, г. Архангельск, Покорителей космоса 67, 123
17	375	100050, г. Москва, Воздвиженка 17, 43
18	376	150015, г. Ярославль, Волкова 53, 45
19	377	170034, г. Астрахань, Лермонтова 66, 88
20	378	120007, г. Санкт-Петербург, Средний пр-кт 30, 2
21	379	Адрес не указан
22	380	120005, г. Санкт-Петербург, Детская 12, 26
23	381	150015, г. Ярославль, Волкова 53, 45
24	382	Адрес не указан
25	383	Адрес не указан
26	384	Адрес не указан
27	385	100050, г. Москва, Воздвиженка 17, 43

Необходимо по установленной схеме отобрать посетителей Web-сайта фирмы для рассылки им каталогов новой продукции.

Для решения задачи используем режим работы «Выборка». Значения параметров, установленных в одноименном диалоговом окне, представлены на рис. 3.2, а сформированная выигрышная выборка – в табл. 3.2.

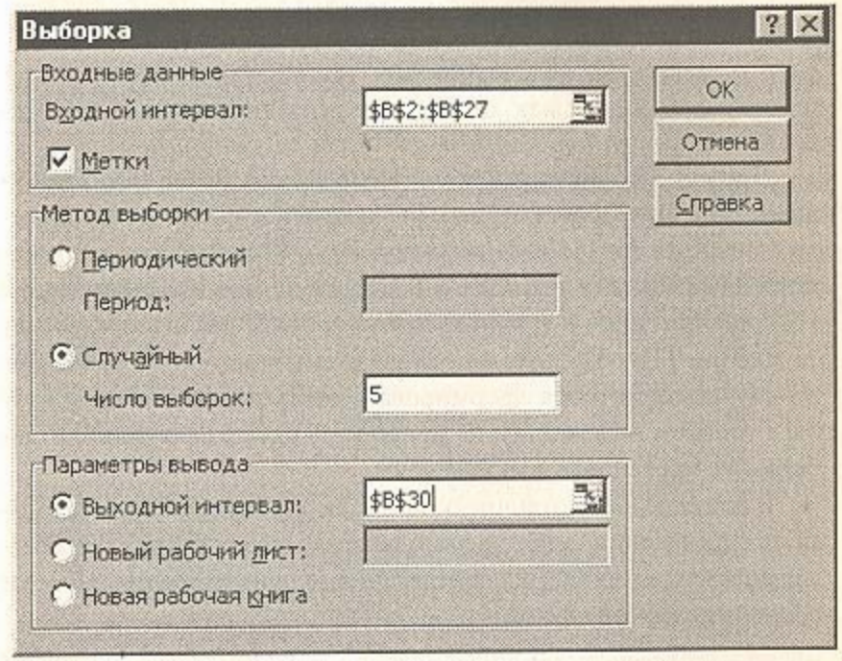

Рис. 3.2

Как видно из табл. 3.2, за последнюю неделю выигрышными оказались 362-е, 365-е, 379-е, 382-е и 385-е посещения Web-сайта фирмы. Причем 362-е и 385-е посещения произвел один и тот же клиент, поэтому в его адрес (100050, г. Москва, Воздвиженка 17, 43) будет выслано два каталога; 365-е посещение оказалось выигрышным для клиента из Тамбова (в его адрес будет выслан один каталог); 379-е и 382-е посещения

Таблица 3.2

	B
30	362
31	379
32	382
33	365
34	385

хотя и оказались выигрышными, по ним рассылка не будет производиться, так как клиенты не указали свои почтовые адреса.

Пример 3.2. Предприятием «Импульс» за месяц было выпущено 1500 приборов, которым были присвоены заводские номера с 10001-го по 11500-й включительно. Все приборы выпускаются на основании технической документации, в соответствии с которой дисперсия чувствительности приборов не превышает 25 мкВ^2/м^2. Необходимо на основе схемы повторного собственно-случайного отбора сформировать контрольную выборку, чтобы с уровнем надежности не менее 95 % предельная ошибка выборки не превышала 3 мкВ/м.

В примере 3.2, в отличие от примера 3.1, важным является момент определения необходимого объема выборки, чтобы она была репрезентативной. Для определения величины объема выборки воспользуемся формулой

$$n = \frac{t^2\sigma^2}{\Delta_x^2},$$

подставляя в которую исходные данные задачи, получим

$$n = \frac{t^2\sigma^2}{\Delta_x^2} = \frac{1{,}96^2 \cdot 25}{3^2} = 10{,}69 \approx 11 \text{ (приборов).}$$

Примечание. В расчете необходимого объема выборки используется коэффициент доверия t, для вычисления которого в Microsoft Excel предусмотрена функция СТЬЮДРАСПОБР (*см.* подразд. 6.3.8). Коэффициент доверия t рассчитывается по формуле =СТЬЮДРАСПОБР(0,05;1499), где 0,05 = 1 – 0,95 – требуемый уровень значимости, 1499 = 1500 – 1 – число степеней свободы.

Таким образом, минимально допустимый объем выборки составляет 11 приборов. При меньшем объеме выборка не будет репрезентативной.

Последующая технология решения задачи аналогична технологии решения задачи в примере 3.1. При этом в поле *Число выборок* вводится рассчитанное значение необходимого объема выборки $n = 11$.

Для быстрого ввода исходных данных (объем генеральной совокупности составляет все же 1500 ед.!) рекомендуем воспользоваться таким техническим приемом, как копирование ячеек с помощью правой клавиши мыши с последующей установкой через контекстное меню *арифметической прогрессии с шагом 1.*

Результатом решения задачи явилась выборка из 11 приборов с заводскими номерами: 10509, 10544, 10769, 10866, 10889, 10902, 10931, 11003, 11087, 11330, 11357. Так как в выборке номера приборов не повторяются, то каждый прибор подвергается проверке только один раз.

Кроме возможности формирования выборки на основе схемы повторного собственно-случайного отбора режим «Выборка» позволяет формировать выборочную совокупность из периодических данных. Порядок формирования такой выборки рассмотрим на следующем примере.

Пример 3.3. В табл. 3.3 приведена сравнительная динамика платных услуг населению Ярославской обл. в 1997 и 1998 гг. (в сопоставимых ценах) [2]. На основе представленной информации необходимо построить графики динамики по квартальным данным.

Для построения графиков необходимо предварительно сформировать таблицу квартальных данных. Это легко делается

Таблица 3.3

	E	F	G
4	Сравнительная динамика объема платных услуг населению Ярославской области в 1997 и 1998 гг. (в сопоставимых ценах), млн руб.		
5		**1997**	**1998**
6	Январь	173,0	146,8
7	Февраль	175,3	155,7
8	Март	186,2	166,5
9	**I квартал**	**534,5**	**469,0**
10	Апрель	186,1	162,3
11	Май	184,9	157,5
12	Июнь	207,7	178,2
13	**II квартал**	**578,7**	**498,0**
14	Июль	239,9	209,4
15	Август	225,9	199,5
16	Сентябрь	218,7	195,5
17	**III квартал**	**684,5**	**604,4**
18	Октябрь	213,9	193,8
19	Ноябрь	232,0	216,0
20	Декабрь	216,1	204,2
21	**IV квартал**	**662,0**	**614,0**

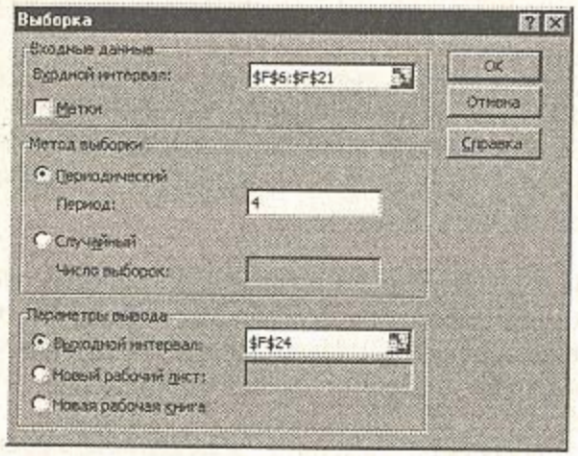

Рис. 3.3

Таблица 3.4

	E	F	G
23		1997	1998
24	I квартал	534,5	469,0
25	II квартал	578,7	498,0
26	III квартал	684,5	604,4
27	IV квартал	662,0	614,0

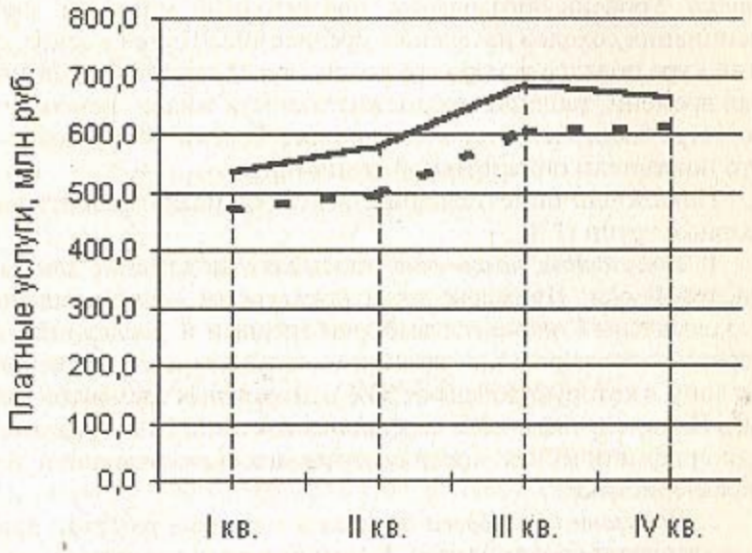

Рис. 3.4

в режиме работы «Выборка». Значения параметров, установленных в одноименном диалоговом окне для первого диапазона данных, показаны на рис. 3.3. Сформированные выборки приведены в табл. 3.4, а построенные с помощью мастера диаграмм графики – на рис. 3.4 (сплошной линией – график динамики объема платных услуг за 1997 г., пунктирной – за 1998 г.).

ГЛАВА 4
Описательная статистика

4.1. Краткие сведения из теории статистики

Статистическая информация представляется совокупностью данных, для характеристики которых используются разнообразные показатели, называемые показателями *описательной статистики*. Уровень образования, прожиточный минимум, дифференциация доходов населения, среднее число детей в семье, средний курс доллара и мера его колебания за определенный интервал времени, таблицы продолжительности жизни, наиболее часто встречающийся счет в чемпионате России по футболу – все это показатели описательной статистики.

Показатели описательной статистики можно разбить на несколько групп [13].

1. *Показатели положения* описывают положение данных на числовой оси. Примеры таких показателей – минимальный и максимальный элементы выборки (первый и последний члены вариационного ряда), верхний и нижний квартили (ограничивают зону, в которую попадают 50% центральных элементов выборки). Наконец, сведения о середине совокупности могут дать средняя арифметическая, средняя гармоническая, медиана и другие характеристики.

2. *Показатели разброса* описывают степень разброса данных относительно своего центра. К ним в первую очередь относятся: дисперсия, стандартное отклонение, размах выборки (разность между максимальным и минимальным элементами), межквартильный размах (разность между верхней и нижней квартилью), эксцесс и т. п. Эти показатели определяют, насколько кучно основная масса данных группируется около центра.

3. *Показатели асимметрии* характеризуют симметрию распределения данных около своего центра. К ним можно отнести коэффициент асимметрии, положение медианы относительно среднего и т. п.

4. *Показатели, описывающие закон распределения*, дают представление о законе распределения данных. Сюда относятся таблицы частот, таблицы частостей, полигоны, кумуляты, гистограммы (*см.* подразд. 2.1).

На практике чаще всего используются следующие показатели: средняя арифметическая, медиана, дисперсия, стандартное отклонение. Однако для получения более точных и достоверных выводов необходимо учитывать и другие из перечисленных выше характеристик, а также обращать внимание на условия получения выборочных совокупностей. Наличие выбросов, т. е. грубых ошибочных наблюдений, может не только сильно исказить значения выборочных показателей (выборочного среднего, дисперсии, стандартного отклонения и т. д.), но и привести ко многим другим ошибочным выводам.

4.2. Справочная информация по технологии работы

Режим «Описательная статистика» служит для генерации одномерного статистического отчета по основным показателям *положения, разброса* и *асимметрии выборочной совокупности*.

В диалоговом окне данного режима (рис. 4.1) задаются следующие параметры:

1. *Входной интервал* – *см.* подразд. 1.1.2.

2. *Группирование* – *см.* подразд. 1.1.2.

3. *Метки в первой строке/Метки в первом столбце* – *см.* подразд. 1.1.2.

4. *Выходной интервал/Новый рабочий лист/Новая рабочая книга* – *см.* подразд. 1.1.2.

5. *Итоговая статистика* – установите в активное состояние, если в выходном диапазоне необходимо получить по одному полю для каждого из следующих показателей описательной статистики: средняя арифметическая выборки ($\bar{x}$), средняя ошибка выборки ($\mu_{\bar{x}}$), медиана (*Me*), мода (*Mo*), оценка стандартного отклонения по выборке (σ), оценка дисперсии по выборке (D), оценка эксцесса по выборке (E_k), оценка коэффициента асимметрии по выборке (A_s), размах вариации выборки (R), минимальный и макси-

мальный элементы выборки, сумма элементов выборки, количество элементов в выборке, k-й наибольший и k-й наименьший элементы выборки, предельная ошибка выборки ($\Delta_{\bar{x}}$).

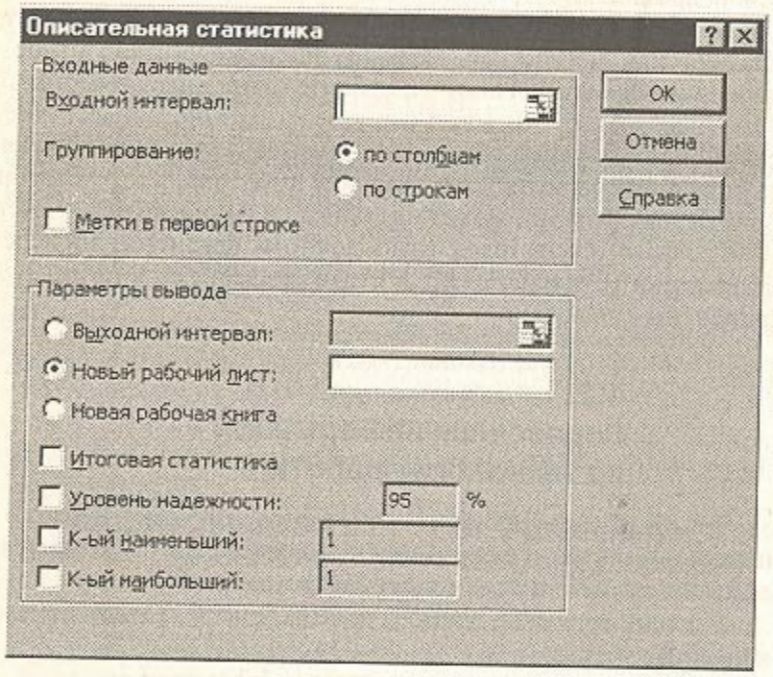

Рис. 4.1

6. *Уровень надежности* – установите в активное состояние, если в выходную таблицу необходимо включить строку для предельной ошибки выборки ($\Delta_{\bar{x}}$) при установленном уровне надежности. В поле, расположенном напротив флажка, введите требуемое значение уровня надежности (например, значение уровня надежности 95 % равносильно доверительной вероятности $\gamma = 0{,}95$ или уровню значимости $\alpha = 0{,}05$).

7. *К-й наибольший* – установите в активное состояние, если в выходную таблицу необходимо включить строку для k-го наибольшего (начиная с максимума x_{max}) значения элемента выборки. В поле, расположенное напротив флажка, введите число k. Если $k = 1$, то строка будет содержать максимальное значение элемента выборки.

8. *К-й наименьший* – установите в активное состояние, если в выходную таблицу необходимо включить строку для k-го наименьшего (начиная с минимума x_{min}) значения элемента выборки. В поле, расположенное напротив флажка, введите число k. Если $k = 1$, то строка будет содержать минимальное значение элемента выборки.

Пример 4.1. Стоимость набора из 25 продуктов питания по некоторым городам центрального региона России по состоянию на декабрь 1998 г. приведена в табл. 4.1 [2], сформированной на рабочем листе Microsoft Excel.

Таблица 4.1

	A	B
1	Стоимость набора из 25 продуктов питания по некоторым городам центрального региона России в декабре 1998 г., руб.	
2	Владимир	389,04
3	Вологда	417,78
4	Иваново	394,00
5	Кострома	371,96
6	Москва	525,96
7	Нижний Новгород	405,12
8	Рязань	419,52
9	Тверь	401,93
10	Ярославль	418,97

Необходимо рассчитать основные показатели описательной статистики и сделать соответствующие выводы.

Для решения задачи используем режим работы «Описательная статистика». Значения параметров, установленных в одноименном диалоговом окне, представлены на рис. 4.2, а показатели, рассчитанные в данном режиме, – в табл. 4.2 (результаты округлены до двух значащих цифр).

Таблица 4.2

	A	B
14	Столбец 1	
15		
16	Среднее	416,03
17	Стандартная ошибка	14,71
18	Медиана	405,12
19	Мода	#Н/Д
20	Стандартное отклонение	44,13
21	Дисперсия выборки	1947,78
22	Эксцесс	6,06
23	Асимметричность	2,26
24	Интервал	154,00
25	Минимум	371,96
26	Максимум	525,96
27	Сумма	3744,28
28	Счет	9,00
29	Наибольший (1)	525,96
30	Наименьший (1)	371,96
31	Уровень надежности (95,0%)	33,92

На основании проведенного выборочного обследования (*см.* табл. 4.1) и рассчитанных по данной выборке показателей описательной статистики (*см.* табл. 4.2) с уровнем надежности 95% можно предположить, что средняя стоимость набора из 25 продуктов питания в целом по всем городам центрального региона России в декабре 1998 г. находилась в пределах от 382,11 до 449,95 руб.

Поясним, на основании каких показателей описательной статистики был сформулирован соответствующий вывод. Такими показателями являются: средняя арифметическая выборки $\bar{x}$ (показатель *Среднее* в табл. 4.2) и предельная ошибка выборки $\Delta_{\bar{x}}$(показатель *Уровень надежности* (95,0%) в табл. 4.2). Из выражения для доверительного интервала

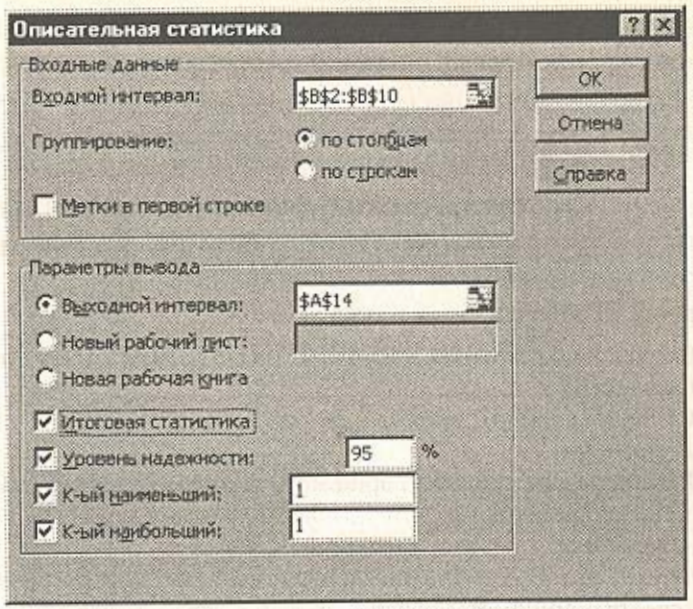

Рис. 4.2

$$\bar{x} - \Delta_{\bar{x}} \le \bar{x} \le \bar{x} + \Delta_{\bar{x}}$$

находим: 416,03 – 33 ,92 = 382,11 – *левая граница*; 416,03 + 33,92 = = 449,95 – *правая граница*.

Коэффициент вариации

$$v = \frac{\sigma}{\bar{x}} \cdot 100\% = \frac{44{,}13}{416{,}03} 100\% \approx 10{,}6\%$$

существенно меньше 40 %, что свидетельствует о небольшой колеблемости признака в исследованной выборочной совокупности. Надежность средней в выборке подтверждается также и ее незначительным отклонением от медианы: 416,03 – 405,12 = 10,91. Значительные положительные значения коэффициентов асимметрии (A_s) и эксцесса (E_k) позволяют говорить о том, что данное эмпирическое распределение существенно отличается от нормального, имеет правостороннюю асимметрию и характеризуется скоплением членов ряда в центре распределения.

Математико-статистическая интерпретация полученных результатов рассмотрена в описании соответствующих статистических функций.

4.3. Статистические функции, связанные с режимом «Описательная статистика»

Функция СРЗНАЧ

См. также СРЗНАЧА, УРЕЗСРЕДНЕЕ, СРГАРМ, СРГЕОМ.

Синтаксис:

СРЗНАЧ (число1; число2; ...)

Результат:

Рассчитывает среднюю арифметическую значений, заданных в списке аргументов.

Аргументы:

число1, число2, ...: от 1 до 30 аргументов, для которых вычисляется средняя арифметическая.

Замечания:

- аргументы должны быть числами или именами, массивами или ссылками, содержащими числа;
- если аргумент, который является массивом или ссылкой, содержит текстовые, логические значения или пустые ячейки, то такие значения игнорируются; однако ячейки, содержащие нулевые значения, учитываются;
- вычисляя средние значения ячеек, следует учитывать различие между пустыми ячейками и ячейками, содержащими нулевые значения, особенно если не установлен флажок *Нулевые Значения* на вкладке ***Вид*** в диалоговом окне **Параметры**. Пустые ячейки не учитываются, но нулевые ячейки учитываются. Чтобы открыть диалоговое окно **Параметры**, выберите команду **Параметры...** в меню ***Сервис***.

Математико-статистическая интерпретация:

Средняя арифметическая является наиболее распространенным видом средних величин. В зависимости от характера имеющихся данных средняя арифметическая может быть *невзвешенной* (*простой*) и *взвешенной*. Функция СРЗНАЧ рассчитывает значение *невзвешенной* средней арифметической по формуле

$$\overline{x} = \frac{\sum x_i}{n}.$$

Рассмотрим использование функции СРЗНАЧ для расчета среднего объема индивидуального жилищного строительства по районам Ярославской области в 1998 г. (табл. 4.3) [2].

Таблица 4.3

	B	C
38	Объем индивидуального жилищного строительства по районам Ярославской области в 1998 г.	
39	Районы	Площадь, м^2
40	Большесельский	718
41	Борисоглебский	1319
42	Брейтовский	632
43	Гаврилов-Ямский	919
44	Даниловский	1321
45	Любимский	437
46	Мышкинский	218
47	Некоузский	206
48	Некрасовский	2121
49	Первомайский	457
50	Переславский	8872
51	Пошехонский	3011
52	Ростовский	1363
53	Рыбинский	1389
54	Тутаевский	730
55	Угличский	4728
56	Ярославский	3439
57	Средний объем строительства	**1875**

Ячейка C57 содержит формулу =СРЗНАЧ(C40:C56), по которой рассчитывается средний объем индивидуального жилищного строительства.

Однако на практике все же наиболее часто приходится иметь дело со *взвешенной* средней арифметической, которая рассчитывается по формуле

$$\bar{x} = \frac{\sum x_i f_i}{\sum f_i}.$$

Взвешенная средняя арифметическая используется при расчете индексов Стендарда и Пура (Standard and Poor's 500 Stock Index), ROS-INDEX и др.

В явном виде функция для расчета взвешенной средней арифметической не представлена в Microsoft Excel, но ее можно легко получить комбинацией других функций. Рассмотрим, как рассчитывается средний курс продажи долларов США по итогам торгов на российских валютных биржах (табл. 4.4 [12]).

Таблица 4.4

	B	C	D
2	Итоги торгов на российских валютных биржах 06.02.95 г.		
3	Валютные биржи	Объем продаж, млн долл.	Курс, руб./долл.
4			
5	Московская межбанковская	72,99	4133
6	Санкт-Петербургская	8,40	4165
7	Сибирская межбанковская	3,97	4126
8	Уральская региональная	25,69	4130
9	Азиатско-Тихоокеанская межбанковская	3,50	4115
10	Ростовская межбанковская	0,64	4127
11	Нижегородская валютно-фондовая	0,02	4133
12	Средний курс продажи долларов США на 06.02.95		4133,8

Ячейка D12 содержит формулу =СУММПРОИЗВ(C5:C11; D5:D11)//СУММ(C5:C11), по которой рассчитывается средневзвешенный курс доллара США по проведенным торгам.

♦ В примере 4.1 значение средней арифметической (показатель *Среднее* в табл. 4.2) рассчитывается формулой =СРЗНАЧ(B2:B10).

Функция МЕДИАНА

См. также МОДА, КВАРТИЛЬ, ПЕРСЕНТИЛЬ.

Синтаксис:

МЕДИАНА (число1; число2; ...)

Результат:

Рассчитывает медиану заданных аргументов.

Аргументы:

число1, число2, ...: от 1 до 30 аргументов, для которых определяется медиана.

Замечания:

• аргументы должны быть числами или именами, массивами или ссылками, содержащими числа;

• если аргумент, который является ссылкой, содержит пустые ячейки, текстовые или логические значения, то такие значения игнорируются; однако ячейки, которые содержат нулевые значения, учитываются.

Математико-статистическая интерпретация:

Медианой (*Me*) называется значение признака, приходящееся на середину *ранжированной (упорядоченной)* совокупности.

Для ранжированного ряда с нечетным числом элементов медианой является варианта, расположенная в центре ряда. Так, данные из табл. 4.5 после ранжировки в порядке возрастания будут представлять последовательность (200, 236, 250, 305, 337). Медианой для данного ряда является третья варианта – 250 костюмов.

Функция МЕДИАНА не требует предварительной ранжировки данных, она проводит ее автоматически. Если в ячейку C9 поместить формулу =МЕДИАНА(C4:C8), то она рассчитает значение 250.

Для ранжированного ряда с четным числом элементов медианой будет средняя арифметическая из двух смежных вариант. Так, функция =МЕДИАНА(200;236;250;305;337;220) рассчитает значение медианы 243 = (236 + 250)/2.

Таблица 4.5

	B	C
2	*Спрос на спортивные костюмы в фирме «Чемпион» (за 2000 г.)*	
3	Производитель костюмов	Число купленных костюмов
4	Diadora	236
5	Adidas	200
6	Reebok	337
7	Nike	250
8	Umbro	305
9	Медиана	**250**

Главное свойство медианы заключается в том, что сумма абсолютных отклонений членов ряда от медианы есть величина наименьшая: $\Sigma|x_i - Me| = \min$.

В отличие от дискретных вариационных рядов определение медианы по интервальным рядам требует проведения определенных расчетов. Так как медиана делит численность ряда пополам, то, следовательно, она будет там, где накопленная частота составляет половину или больше половины всей суммы частот, а предыдущая накопленная частота меньше половины численности совокупности.

Если предполагать, что внутри медианного интервала нарастание или убывание изучаемого признака происходит по прямой равномерно, то формула медианы в интервальном ряду распределения будет иметь следующий вид:

$$Me = x_0 + i\frac{0{,}5\sum f_i - S_{Me-1}}{f_{Me}},$$

где x_0 – нижняя граница медианного интервала;
i – величина медианного интервала;
f_{Me} – частота медианного интервала;
S_{Me-1} – накопленная частота интервала, предшествующего медианному.

В табл. 4.6 медианным интервалом величины научного стажа сотрудников научно-исследовательского центра будет интервал 8–10 лет, а медианой продолжительности стажа – 8,13 лет.

Таблица 4.6

	B	C	D
2	Научный стаж сотрудников НИЦ, лет	Число сотрудников, f_i	Накопленная частота, S_i
3	До 4	14	14
4	4–6	33	47
5	6–8	30	77
6	8–10	45	122
7	10–12	21	143
8	Свыше 12	17	160
9	Итого	160	
10	50% числа сотрудников	80	
11	Смещение на max ≤ N/2	3	
12	Значение max ≤ N/2	77	
13	Смещение на медианный интервал	4	
14	Частота медианного интервала	45	
15	Медианный интервал	8–10	
16	Нижняя граница медианного интервала	8	
17	Значение накопленной частоты предшествующего интервала	77	
18	Медиана продолжительности стажа	**8,13**	

Содержимое ячеек в табл. 4.6:

• в ячейках D3:D8 вычисляются накопленные частоты (например, ячейка D5 содержит формулу =D4+C5);

• ячейка C9 содержит формулу =СУММ(C3:C8) – рассчитывается численность совокупности (число сотрудников);

• ячейка C10 содержит формулу =C9/2 – определяется половина численности совокупности (50 % числа сотрудников);

• ячейка C11 содержит формулу =ПОИСКПОЗ(C10;D3:D8;1) – в массиве D3:D8 определяется номер позиции числа, которое является наибольшим среди чисел меньших или равных середины интервала, т. е. числа 80;

• ячейка C12 содержит формулу =ИНДЕКС(D3:D8;C11;1) – из массива D3:D8 извлекается число, удовлетворяющее условиям поиска, сформированным в ячейке C11;

• ячейка C13 содержит формулу =ЕСЛИ(C10=C12;C11;C11+1) – рассчитывается смещение на медианный интервал;

• ячейка C14 содержит формулу =ИНДЕКС(C3:C8;C13;1)– отображается значение частоты медианного интервала;

• ячейка C15 содержит формулу =ИНДЕКС(B3:B8;C13;1) – в массиве B3:B8 находится медианный интервал;

• ячейка C16 содержит формулу =ЛЕВСИМВ(C15;1) – отображается нижняя граница медианного интервала;

• ячейка C17 содержит формулу =ИНДЕКС(D3:D8;C13–1;1) – находится значение накопленной частоты интервала, предшествующего медианному;

• ячейка C18 содержит формулу =C16+2*((C9/2–C17)/C14) – рассчитывается медиана продолжительности стажа.

Безусловно, из приведенных формул можно составить одну интегрированную формулу (*см.*, например, описание функции МОДА). Однако с целью более быстрого составления и поиска возможных ошибок рекомендуется сложные формулы составлять по частям.

♦ В примере 4.1 значение медианы (показатель *Медиана* в табл. 4.2) рассчитывается по формуле =МЕДИАНА(B2:B10).

Функция МОДА

См. также МЕДИАНА.

Синтаксис:

МОДА (число1; число2; ...)

Результат:

Отображает наиболее часто встречающееся значение в интервале данных.

Аргументы:

число1, число2, ...: от 1 до 30 аргументов, для которых вычисляется мода.

Замечания:

• аргументы должны быть числами, именами, массивами или ссылками, которые содержат числа;

• если аргумент, который является массивом или ссылкой, содержит текстовые, логические значения или пустые ячейки, то такие значения игнорируются, однако ячейки, содержащие нулевые значения, учитываются;

• если множество данных не содержит одинаковых данных, то функция МОДА помещает в ячейку значение ошибки #Н/Д.

Математико-статистическая интерпретация:

Модой (*Мо*) называется чаще всего встречающаяся варианта или то значение признака, которое соответствует максимальной точке теоретической кривой распределения.

Мода широко используется в коммерческой практике при изучении покупательского спроса (при определении «ходовых» размеров одежды и обуви, наиболее употребляемых продуктов и т. п.). В дискретном ряду мода – это варианта с наибольшей частотой. По данным, приведенным в табл. 4.7, можно судить, что наибольшим спросом пользуются спортивные костюмы 50 размера, соответственно он и является модальным.

Однако, если использовать функцию МОДА, то в ячейку C10 она поместит значение ошибки #Н/Д. Это объясняется тем, что функция МОДА находит наиболее часто встречающееся значение *в интервале* данных. Так, если в ячейку C10 ввести формулу =МОДА(B4:B8;52), то она поместит в ячейку значение 52.

Для получения модального (максимального) значения по данным табл. 4.7 в ячейку C9 введена формула =МАКС(C4:C8), а в ячейку C10 – ИНДЕКС(B4:B8;ПОИСКПОЗ(C9;C4:C8)).

В отличие от дискретного вариационного ряда определение моды *по интервальному ряду* требует проведения расчетов по формуле

Таблица 4.7

	B	C
2	*Спрос на спортивные костюмы «Reebok» в фирме «Чемпион» (за 2000 г.)*	
3	Размер костюма	Число купленных костюмов
4	46	57
5	48	48
6	50	95
7	52	60
8	54	77
9	Наибольший спрос	95
10	Мода	**50**

$$Mo = x_0 + i\frac{(f_{Mo} - f_{Mo-1})}{(f_{Mo} - f_{Mo-1}) + (f_{Mo} - f_{Mo+1})},$$

где x_0 – нижняя граница модального интервала;
i – величина модального интервала;
f_{Mo} – частота модального интервала;
f_{Mo-1} – частота интервала, предшествующего модальному;
f_{Mo+1} – частота интервала, следующего за модальным.

В табл. 4.8 модальным интервалом продолжительности стажа сотрудников научно-исследовательского центра (НИЦ) является интервал 8–10 лет, а модой продолжительности стажа – 8,77 лет.

Ячейка С9 содержит формулу =ЛЕВСИМВ(ИНДЕКС(B3:B8; ПОИСКПОЗ(МАКС(C3:C8);C3:C8;0);1);1)+2*((МАКС(C3:C8)–ИНДЕКС(C3:C8;ПОИСКПОЗ(МАКС(C3:C8);C3:C8;0)–1;1))/((МАКС(C3:C8)–ИНДЕКС(C3:C8;ПОИСКПОЗ(МАКС (C3:C8); C3:C8;0)–1;1))+(МАКС(C3:C8)–ИНДЕКС(C3:C8;ПОИСКПОЗ(МАКС (C3:C8); C3:C8;0)+1;1)))).

Безусловно, представленная формула слишком громоздка и непонятна. Она приведена только лишь для демонстрации того факта, что, работая с Microsoft Excel, можно обойтись без промежуточных вычислений на рабочем листе и заключить все расчеты в одну формулу. Но даже если в этом и есть необходимость, советуем подобные формулы разрабатывать по частям, что предотвратит от ошибок и сэкономит время.

С целью пояснения представленной формулы рассмотрим табл. 4.9, функционально адекватную табл. 4.8.

Таблица 4.8

	B	C
2	Научный стаж сотрудников НИЦ, лет	Число сотрудников, f_i
3	До 4	14
4	4–6	33
5	6–8	30
6	8–10	45
7	10–12	21
8	Свыше 12	17
9	Наиболее часто встречающийся стаж	8,77

Содержимое ячеек в табл. 4.9:

- ячейка С9 содержит формулу =МАКС(C3:C8) – рассчитывается модальная численность сотрудников;
- ячейка С10 содержит формулу =ПОИСКПОЗ(C9;C3:C8;0) – в массиве C3:C8 вычисляется смещение на модальное значение;
- ячейка С11 содержит формулу =ИНДЕКС(B3:B8;C10;1) – в массиве B3:B8 находится модальный интервал стаж;
- ячейка С12 содержит формулу =ЛЕВСИМВ(C11;1) – отображается нижняя граница модального интервала стажа;
- ячейка С13 содержит формулу =ИНДЕКС(C3:C8;C10–1;1) – в массиве C3:C8 находится число сотрудников с предшествующим стажем;
- ячейка С14 содержит формулу =ИНДЕКС(C3:C8;C10+1;1) – в массиве C3:C8 находится число сотрудников с последующим стажем;
- ячейка С15 содержит формулу =C12+2*((C9–C13)/((C9–C13)+(C9–C14))) – рассчитывается мода продолжительности стажа.

♦ В примере 4.1 значение моды (показатель *Мода* в табл. 4.2) рассчитывается по формуле =МОДА(B2:B10).

Таблица 4.9

	B	C
2	Научный стаж сотрудников НИЦ, лет	Число сотрудников, f_i
3	До 4	14
4	4–6	33
5	6–8	30
6	8–10	45
7	10–12	21
8	Свыше 12	17
9	Модальная численность сотрудников	45
10	Смещение в столбце на модальное значение	4
11	Модальный интервал стажа	8–10
12	Нижняя граница модального интервала	8
13	Число сотрудников с предшествующим стажем	30
14	Число сотрудников с последующим стажем	21
15	Мода продолжительности стажа	**8,77**

Функция СТАНДОТКЛОН

См. также ДИСП, КВАДРОТКЛ, СРОТКЛ, СТАНДОТКЛОНА, СТАНДОТКЛОНП.

Синтаксис:

СТАНДОТКЛОН (число1; число2; ...)

Результат:

Оценивает генеральное стандартное отклонение по *выборке*.

Аргументы:

число1, число2, ...: от 1 до 30 аргументов, соответствующих выборке из генеральной совокупности.

Замечания:

• функция СТАНДОТКЛОН предполагает, что аргументы являются выборкой из генеральной совокупности. Если данные представляют всю генеральную совокупность, то стандартное отклонение следует вычислять с помощью функции СТАНДОТКЛОНП;

• логические значения, такие, как ИСТИНА или ЛОЖЬ, а также текст игнорируются. Если текстовые и логические значения игнорироваться не должны, следует использовать функцию СТАНДОТКЛОНА.

Математико-статистическая интерпретация:

См. описание функций ДИСП и СТАНДОТКЛОНП.

Внимание! Функция СТАНДОТКЛОН рассчитывает генеральное стандартное отклонение при условии, что исходные данные образуют *выборочную* совокупность. Если совокупность является *генеральной*, необходимо воспользоваться функцией СТАНДОТКЛОНП.

Используя выборочные данные, приведенные в табл. 4.10, по формуле =СТАНДОТКЛОН(C4:C9) получим стандартное отклонение 94,66 (сравните со значением 86,41, вычисляемым функцией СТАНДОТКЛОНП).

♦ В примере 4.1 значение стандартного отклонения (показатель *Стандартное отклонение* в табл. 4.2) рассчитывается по формуле =СТАНДОТКЛОН(B2:B10).

Таблица 4.10

	B	C
2	Сбор зерна по некоторым хозяйствам района	
3	Название хозяйств	Валовой сбор зерна, ц
4	Петровка	600
5	Ивановка	520
6	Сидоровка	400
7	Пантелеевка	600
8	Сергеевка	500
9	Андреевка	380
10	Стандартное отклонение	**94,66**

В режиме «Описательная статистика» функция СТАНДОТКЛОН совместно с функцией СЧЕТ используется также для определения средней ошибки выборки $\mu_{\bar{x}}$ (показатель *Стандартная ошибка* в табл. 4.2).

Средняя ошибка выборки характеризует стандартное отклонение вариантов выборочной средней от генеральной средней и зависит от колеблемости признака в генеральной совокупности σ, числа отобранных единиц n, а также от способа организации выборки. Средняя ошибка повторной собственно-случайной выборки определяется по формуле

$$\mu_{\bar{x}} = \frac{\sigma}{\sqrt{n}},$$

где σ – оценка генерального стандартного отклонения;
n – объем выборочной совокупности.

♦ В примере 4.1 значение средней ошибки выборки (показатель *Стандартная ошибка* в табл. 4.2) рассчитывается по формуле =B20/КОРЕНЬ(B28),

где в ячейке B20 – значение оценки генерального стандартного отклонения, рассчитываемого по формуле =СТАНДОТКЛОН (B2:B10);
в ячейке B28 – значение объема выборки, рассчитываемого по формуле =СЧЕТ(B2:B10).

Средняя ошибка выборки $\mu_{\bar{x}}$ используется для расчета предельной ошибки выборки $\Delta_{\bar{x}}$ (показатель *Уровень надежности* в табл. 4.2), которая дает возможность выяснить, в каких пределах находится величина генеральной средней.

В математической статистике установлено, что предельная ошибка выборки $\Delta_{\bar{x}}$ связана со средней ошибкой выборки $\mu_{\bar{x}}$ соотношением

$$\Delta_{\bar{x}} = t\mu_{\bar{x}},$$

где t – коэффициент доверия (определяется в зависимости от того, с какой доверительной вероятностью нужно гарантировать результаты выборочного обследования).

В Microsoft Excel коэффициент доверия t рассчитывается через функцию СТЬЮДРАСПОБР (*см.* подразд. 6.3.8), в которой в качестве аргументов задаются уровень значимости α и число степеней свободы k. Уровень значимости α связан с доверительной вероятностью γ (задается в поле *Уровень надежности* диалогового окна **Описательная статистика**, рис. 4.1) выражением $\alpha = 1 - \gamma$. Число степеней свободы k зависит от объема выборки n и связано с ним выражением $k = n - 1$.

♦ В примере 4.1 значение предельной ошибки выборки с уровнем надежности 95% (показатель *Уровень надежности* в табл. 4.2) рассчитывается по формуле =B17*СТЬЮДРАСПОБР(0,05;B28–1),

где в ячейке B17 – значение средней ошибки выборки, рассчитываемое по формуле =B20/КОРЕНЬ(B28));
в ячейке B28 – значение объема выборки, рассчитываемое по формуле =СЧЕТ(B2:B10).

Внимание! В раздел статистических функций Microsoft Excel для вычисления значения предельной ошибки выборки включена также и функция ДОВЕРИТ (*см.* подразд. 6.3.1). Данную функцию можно использовать при сравнительно большом числе единиц выборочной совокупности ($n > 100$), когда расхождение между средней выборки и генеральной средней становится практически несущественным (распределение Стьюдента приближается к нормальному распределению). Для *малых выборок* это расхождение может быть весьма существенным, поэтому для расчета предельной ошибки выборки в этом случае необходимо пользоваться не нормальным распределением (функцией ДОВЕРИТ), а распределением Стьюдента (функцией СТЬЮДРАСПОБР).

Функция ДИСП

См. также ДИСПА, ДИСПР, КВАДРОТКЛ, СРОТКЛ, СТАНДОТКЛОН.

Синтаксис:
ДИСП (число1; число2; ...)

Результат:
Оценивает генеральную дисперсию *по выборке*.

Аргументы:

число1, число2, ...: от 1 до 30 аргументов, соответствующих выборке из генеральной совокупности.

Замечания:

• логические значения, такие, как ИСТИНА или ЛОЖЬ, а также текст игнорируются. Если они не должны игнорироваться, пользуйтесь функцией ДИСПА;

• функция ДИСП предполагает, что аргументы являются выборкой из генеральной совокупности. Если данные представляют всю генеральную совокупность, вычисляйте дисперсию, используя функцию ДИСПР.

Математико-статистическая интерпретация:

См. описание функции ДИСПР.

В связи с тем, что изучаемые статистикой признаки варьируются, обобщающие показатели в выборке могут в той или иной мере отличаться от значений этих характеристик в генеральной совокупности. Причем, чем меньше объем выборки, тем больше вероятность отклонения статистических характеристик от истинных, полученных по генеральной совокупности.

В математической статистике доказывается, что дисперсия выборочной совокупности является состоятельной, но смещенной оценкой генеральной совокупности:

$$M[\sigma^2_{VIB}] = \sigma^2_{GEN}\,\frac{n-1}{n}.$$

Для устранения систематической ошибки и получения несмещенной оценки нужно σ^2_{VIB} умножить на $n/(n-1)$. Тогда при малом числе наблюдений (особенно при $n \le 40-50$) дисперсию σ^2_{GEN} следует вычислять по формуле

$$\sigma^2_{GEN} = \sigma^2_{VIB}\,\frac{n}{n-1}.$$

Поскольку значение $n/(n-1)$ при достаточно больших n близко к 1 (при $n = 100$ значение $n/(n-1) = 1{,}01$, а при $n = 500$ значение $n/(n-1) = 1{,}002$ и т. д.), можно приближенно считать, что выборочная дисперсия равна генеральной дисперсии, т. е. $\sigma^2_{GEN} \approx \sigma^2_{VIB}$.

Используя выборочные данные, приведенные в табл. 4.10, по формуле =ДИСП(C4:C9) получим оценку генеральной дисперсии 8960.

Внимание! Функция ДИСП рассчитывает генеральную дисперсию при условии, что исходные данные образуют *выборочную* совокупность. В случае если совокупность является *генеральной*, необходимо воспользоваться функцией ДИСПР. Так, предположив, что исходные данные в ячейках C4:C9 образуют *генеральную* совокупность, и применив функцию ДИСПР, получим значение генеральной дисперсии, равное 7466,67.

♦ В примере 4.1 значение дисперсии (показатель *Дисперсия выборки* в табл. 4.2) рассчитывается по формуле =ДИСП(B2:B10).

Функция ЭКСЦЕСС

См. также СКОС.

Синтаксис:

ЭКСЦЕСС (число1; число2; ...)

Результат:

Оценивает эксцесс по *выборке*.

Аргументы:

число1, число2, ...: от 1 до 30 аргументов, для которых вычисляется эксцесс.

Замечания:

• аргументы должны быть числами или именами, массивами или ссылками, содержащими числа;

• если аргумент, который является массивом или ссылкой, содержит текстовые, логические значения или пустые ячейки, то такие значения игнорируются, однако ячейки с нулевыми значениями учитываются;

• если задано менее четырех точек данных или если стандартное отклонение выборки равняется нулю, то функция ЭКСЦЕСС помещает в ячейку значение ошибки #ДЕЛ/0!.

Математико-статистическая интерпретация:

Эксцесс характеризует так называемую «крутость», т. е. островершинность или плосковершинность распределения. Он может быть рассчитан для любых распределений, но в большинстве слу-

чаев вычисляется только для симметричных. Это объясняется тем, что за исходную принята кривая нормального распределения ($E_k = 0$), относительно вершины которой и определяется выпад вверх или вниз вершины эмпирического распределения. Функция ЭКСЦЕСС рассчитывает значение эксцесса как для симметричных, так и для асимметричных распределений.

Наиболее точным и распространенным является определение эксцесса, основанное на расчете центрального момента 4-го порядка:

$$E_k = \frac{\mu_4}{\sigma^4} - 3.$$

Применение данной формулы дает возможность вычислить значение эксцесса в *генеральной* совокупности. При этом если $E_k > 0$, распределение островершинное (рис. 4.3 *а*), если $E_k < 0$ – плосковершинное (рис. 4.3 *б*).

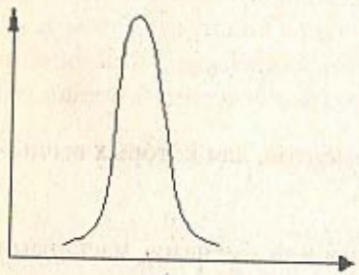

Рис 4.3 *а*

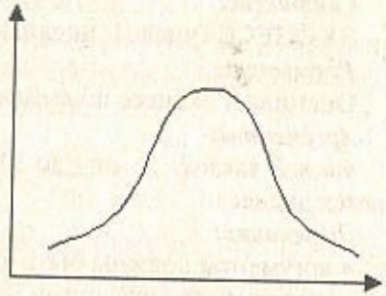

Рис 4.3 *б*

Необходимо отметить, что функция ЭКСЦЕСС определяет значение эксцесса по *выборочной* совокупности, поэтому в ней реализована формула

$$E_k = \left(\frac{n(n+1)}{(n-1)(n-2)(n-3)} \sum \left(\frac{x_i - \bar{x}}{\sigma} \right)^4 \right) - \frac{3(n-1)^2}{(n-2)(n-3)},$$

где n – объем выборки.

Рассмотрим расчет эксцесса по выборочным данным, представленным в табл. 4.11.

Таблица 4.11

	B	C	D
2	№ п/п	Фамилия преподавателя	Педагогический стаж преподавателя вуза, лет
3	1	Орлов	15
4	2	Грачев	8
5	3	Петухов	10
6	4	Голубев	7
7	5	Курочкин	5
8	6	Соловьев	10
9	7	Синицын	5
10	8	Воробьев	2
11	9	Ласточкин	10
12	Эксцесс		**0,41**

Ячейка D12 содержит формулу =ЭКСЦЕСС(D3:D11).

Если данные образуют не выборочную, а генеральную совокупность, то эксцесс необходимо рассчитывать по стандартной формуле через центральный момент 4-го порядка и стандартное отклонение (табл. 4.12).

Содержимое ячеек в табл. 4.12:

- ячейка C14 содержит формулу =СУММ(C4:C13) – рассчитывается общее количество абитуриентов;
- ячейка C15 содержит формулу {=СУММПРОИЗВ(B4:B13; C4:C13)/C14} – определяется средний балл сдачи экзаменов;
- ячейка C16 содержит формулу {=СУММПРОИЗВ(СТЕПЕНЬ(B4:B13–C15;4);C4:C13)/C14} – вычисляется центральный момент 4-го порядка;
- ячейка C17 содержит формулу {=КОРЕНЬ(СУММПРОИЗВ(СТЕПЕНЬ(B4:B13–C15;2);C4:C13)/C14)} – рассчитывается стандартное отклонение;

Таблица 4.12

	B	C
2	Результаты сдачи вступительных экзаменов по математике	
3	Оценка (в баллах по 10-балльной шкале)	Количество абитуриентов
4	1	1
5	2	7
6	3	15
7	4	85
8	5	174
9	6	136
10	7	67
11	8	46
12	9	22
13	10	9
14	Общее количество абитуриентов	562
15	Средний балл	5,71
16	Центральный момент 4-го порядка	18,62
17	Стандартное отклонение	1,54
18	4-я степень стандартного отклонения	5,62
19	Эксцесс	0,31

• ячейка C18 содержит формулу =СТЕПЕНЬ(C17;4) – вычисляется 4-я степень стандартного отклонения;

• ячейка C19 содержит формулу =C16/C18–3 – рассчитывается эксцесс.

В табл. 4.11 и 4.12 эмпирические распределения имеют положительный эксцесс, т. е. они характеризуются скоплением членов ряда в центрах распределений.

Для приблизительного определения значения эксцесса по данным генеральной совокупности (или по данным выборочной

совокупности, имеющей значительный объем) можно также пользоваться упрощенной формулой Линдберга

$$E_k = P - 38{,}29,$$

где P – доля (%) количества вариант, лежащих в интервале, равном половине стандартного отклонения в ту и другую сторону от $\bar{x}$;

38,29 – доля (%) количества вариант, лежащих в интервале, равном половине стандартного отклонения в ту и другую сторону от $\bar{x}$ ряда нормального распределения.

Так, для данных, приведенных в табл. 4.12, эксцесс, рассчитанный по формуле Линдберга, равен 16,87 % (или 0,17).

♦ В примере 4.1 значение эксцесса (показатель *Эксцесс* в табл. 4.2) рассчитывается по формуле =ЭКСЦЕСС(B2:B10).

Функция СКОС

См. также ЭКСЦЕСС.

Синтаксис:

СКОС (число1; число2; ...)

Результат:

Оценивает коэффициент асимметрии по *выборке*.

Аргументы:

число1, число2, ...: от 1 до 30 аргументов, для которых вычисляется коэффициент асимметрии.

Замечания:

• аргументы должны быть числами или именами, массивами или ссылками, содержащими числа;

• если аргумент, который является массивом или ссылкой, содержит текстовые, логические значения или пустые ячейки, то такие значения игнорируются, однако ячейки с нулевыми значениями учитываются;

• если имеется менее трех точек данных или стандартное отклонение равно нулю, то функция СКОС помещает в ячейку значение ошибки #ДЕЛ/0!.

Математико-статистическая интерпретация:

Определение формы кривой является важной задачей, так как статистический материал в обычных условиях дает по определенному признаку характерную, типичную для него кривую распределения. Всякое искажение формы кривой означает нарушение или изменение нормальных условий возникновения статистического материала.

Выяснение общего характера распределения предполагает оценку степени его однородности, а также вычисление показателей асимметрии и эксцесса.

Симметричным является распределение, в котором частоты любых двух вариантов, равноотстоящих в обе стороны от центра распределения, равны между собой.

Для симметричных распределений средняя арифметическая, мода и медиана равны между собой. С учетом этого показатель асимметрии основан на соотношении показателей центра распределения: чем больше разница между $\bar{x}$, *Mo*, *Me*, тем больше асимметрия ряда. При этом если $Mo < Me$, асимметрия правосторонняя, если $Mo > Me$ – асимметрия левосторонняя.

Наиболее точным и часто используемым является показатель, основанный на определении центрального момента 3-го порядка (в симметричном распределении его значение равно нулю):

$$A_S = \frac{\mu_3}{\sigma^3}.$$

Применение данного показателя дает возможность определить величину асимметрии в генеральной совокупности. При этом если $A_s > 0$ – асимметрия правосторонняя (положительная), если $A_s < 0$ – асимметрия левосторонняя (отрицательная) (рис. 4.4*а*, *б*).

Необходимо отметить, что функция СКОС определяет величину асимметрии по *выборочной* совокупности, поэтому в ней реализована формула

$$A_S = \frac{n}{(n-1)(n-2)} \sum \left(\frac{x_i - \bar{x}}{\sigma} \right)^3,$$

где n – объем выборки.

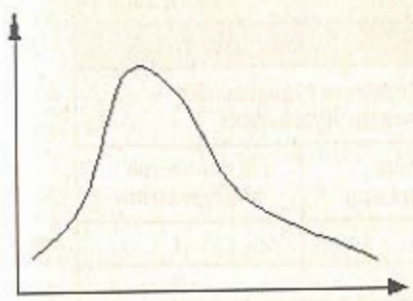

Рис. 4.4*а*

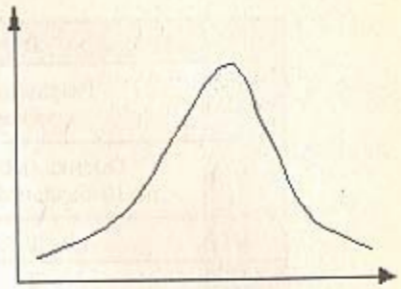

Рис. 4.4*б*

Рассмотрим расчет коэффициента асимметрии по выборочным данным, представленным в табл. 4.13.

Таблица 4.13

	B	C	D
2	№ п/п	Фамилия преподавателя	Педагогический стаж преподавателя вуза, лет
3	1	Орлов	15
4	2	Грачев	8
5	3	Петухов	10
6	4	Голубев	7
7	5	Курочкин	5
8	6	Соловьев	10
9	7	Синицын	5
10	8	Воробьев	2
11	9	Ласточкин	10
12	Коэффициент асимметрии		**0,28**

Ячейка D12 содержит формулу =СКОС(D3:D11).

Если данные образуют не выборочную, а генеральную совокупность, то асимметрию необходимо рассчитывать по стандартной формуле через центральный момент 3-го порядка и стандартное отклонение (табл. 4.14).

Таблица 4.14

	B	C
2	Результаты сдачи вступительных экзаменов по математике	
3	Оценка (в баллах по 10-балльной шкале)	Количество абитуриентов
4	1	1
5	2	7
6	3	15
7	4	85
8	5	174
9	6	136
10	7	67
11	8	46
12	9	22
13	10	9
14	Общее количество абитуриентов	562
15	Средний балл	5,71
16	Центральный момент 3-го порядка	1,66
17	Стандартное отклонение	1,54
18	Куб стандартного отклонения	3,65
19	Коэффициент асимметрии	**0,45**

Содержимое ячеек в табл. 4.14:

• ячейка C14 содержит формулу =СУММ(C4:C13) – вычисляется общее количество абитуриентов;

• ячейка C15 содержит формулу =СУММПРОИЗВ(B4:B13; C4:C13)/C14 – определяется средний балл сдачи экзаменов;

• ячейка C16 содержит формулу =СУММПРОИЗВ(СТЕПЕНЬ (B4:B13–C15;3);C4:C13)/C14 – рассчитывается центральный момент 3-го порядка;

• ячейка C17 содержит формулу =КОРЕНЬ(СУММПРОИЗВ (СТЕПЕНЬ(B4:B13–C15;2);C4:C13)/C14) – вычисляется стандартное отклонение;

• ячейка C18 содержит формулу =СТЕПЕНЬ(C17;3) – рассчитывается 3-я степень стандартного отклонения;

• ячейка C19 содержит формулу =C16/C18 – рассчитывается коэффициент асимметрии.

В табл. 4.13 и 4.14 эмпирические распределения имеют положительную (правостороннюю) асимметрию, т. е. они характеризуются пологим склоном («хвостом») в правой части распределения.

Для приблизительного определения значения показателя асимметрии по данным генеральной совокупности (или по данным выборочной совокупности, имеющей значительный объем) можно также пользоваться упрощенной формулой Линдберга

$$A_S = P - 50,$$

где P – доля (%) количества тех вариант, которые превосходят среднюю арифметическую в общем количестве вариант данного ряда;

50 – доля (%) вариант, превосходящих среднюю арифметическую ряда нормального распределения.

Так, для данных табл. 4.14 показатель асимметрии, рассчитанный по формуле Линдберга, равен 49,82% (или 0,50).

♦ В примере 4.1 значение показателя асимметрии (показатель *Асимметрия* в табл. 4.2) рассчитывается по формуле =СКОС(B2:B10).

Функция МИН

См. также МИНА, МАКС, МАКСА.

Синтаксис:

МИН (число1; число2; ...)

Результат:

Находит наименьшее значение (x_{min}) в множестве данных.

Аргументы:

число1, число2, ...: от 1 до 30 аргументов, среди которых ищется минимальное значение.

Замечания:

• можно задавать аргументы, которые являются числами, пустыми ячейками, логическими значениями или текстовыми представлениями чисел. Аргументы, являющиеся значениями ошибки или текстами, не преобразуемыми в числа, вызывают значения ошибок;

• если аргумент является массивом или ссылкой, то в нем учитываются только числа. Пустые ячейки, логические значения или текст в массиве или ссылке игнорируются. Если логические значения или текст не должны игнорироваться, следует использовать функцию МИНА;

• если аргументы не содержат чисел, то функция МИН помещает в ячейку значение 0.

Математико-статистическая интерпретация:

Функция МИН применяется для нахождения минимального значения признака. Например, для исходных данных (*см.* табл. 4.3) в описании функции СРЗНАЧ формула =МИН (С40:С56) рассчитает значение 206 (Некоузский район).

♦ В примере 4.1 минимальное значение выборки (показатель *Минимум* в табл. 4.2) рассчитывается по формуле =МИН(В2:В10).

Функция МАКС

См. также МАКСА, МИН, МИНА.

Синтаксис:

МАКС (число1; число2; ...)

Результат:

Находит наибольшее значение (x_{max}) в множестве данных.

Аргументы:

число1, число2, ... : от 1 до 30 аргументов, среди которых ищется максимальное значение.

Замечания:

• можно задавать аргументы, которые являются числами, пустыми ячейками, логическими значениями или текстовыми представлениями чисел. Аргументы, которые являются значениями ошибки или текстами, не преобразуемыми в числа, вызывают значения ошибок;

• если аргумент является массивом или ссылкой, то в нем учитываются только числа. Пустые ячейки, логические значения или

текст в массиве или ссылке игнорируются. Если логические значения или текст не должны игнорироваться, следует использовать функцию МАКСА;

• если аргументы не содержат чисел, то функция МАКС помещает в ячейку значение 0.

Математико-статистическая интерпретация:

Функция МАКС применяется для нахождения максимального значения признака. Например, для исходных данных (*см.* табл. 4.3) в описании функции СРЗНАЧ формула =МАКС(С40:С56) рассчитает значение 8872 (Переславский район).

♦ В примере 4.1 максимальное значение выборки (показатель *Максимум* в табл. 4.2) рассчитывается по формуле =МАКС(В2:В10).

В режиме «Описательная статистика» функции МАКС и МИН используются также для определения размаха вариации *R* (показатель *Интервал* в табл. 4.2).

Размах вариации показывает, насколько велико различие между единицами совокупности, имеющими наибольшее и наименьшее значение признака (например, различие между максимальной и минимальной пенсией различных групп населения, нормами выработки у рабочих определенной специальности или квалификации и т. п.). Размах вариации рассчитывают как разность между наибольшим (x_{max}) и наименьшим (x_{min}) значениями варьирующего признака, т. е.

$$R = x_{max} - x_{min}.$$

♦ В примере 4.1 значение размаха вариации выборки (показатель *Интервал* в табл. 4.2) рассчитывается по формуле = В26 – В25,

где в ячейке В25 – минимальное значение выборки, рассчитываемое по формуле =МИН(В2:В10);

в ячейке В26 – максимальное значение выборки, рассчитываемое по формуле =МАКС(В2:В10).

♦ В примере 4.1 суммарное значение элементов выборки (показатель *Сумма* в табл. 4.2) рассчитывается по формуле =СУММ (В2:В10).

Функция СЧЕТ

См. также СЧЕТЗ.

Синтаксис:

СЧЕТ (значение1; значение2; ...)

Результат:

Рассчитывает *количество чисел* в списке аргументов.

Аргументы:

значение1, значение2, ...: от 1 до 30 аргументов, которые могут содержать данные различных типов или ссылаться на них; в подсчете участвуют только числа.

Замечания:

- учитываются аргументы, которые являются числами, пустыми значениями, логическими значениями, датами или текстами, изображающими числа;
- аргументы, являющиеся значениями ошибки или текстами, которые нельзя интерпретировать как числа, игнорируются;
- если аргумент является массивом или ссылкой, то подсчитываются только числа в этом массиве или ссылке. Пустые ячейки, логические значения, тексты и значения ошибок в массиве или ссылке игнорируются.

Математико-статистическая интерпретация:

Функция СЧЕТ используется для получения количества *числовых* ячеек в массивах ячеек.

Для исходных данных (*см.* табл. 4.13) в описании функции СКОС формула =СЧЕТ(D3:D11) рассчитывает значение 9, а формула =СЧЕТ(C3:C11) – значение 0.

♦ В примере 4.1 объем выборочной совокупности (показатель *Счет* в табл. 4.2) рассчитывается по формуле =СЧЕТ(B2:B10).

Функция НАИБОЛЬШИЙ

См. также НАИМЕНЬШИЙ, МАКС.

Синтаксис:

НАИБОЛЬШИЙ (массив; *k*)

Результат:

Находит *k*-е по порядку (начиная с x_{max}) наибольшее значение в множестве данных.

Аргументы:

- *массив:* массив данных, для которых определяется *k*-е наибольшее значение;
- *k*: позиция (начиная с наибольшей) в массиве ячеек данных.

Замечания:

- если массив пуст, то функция НАИБОЛЬШИЙ помещает в ячейку значение ошибки #ЧИСЛО!;
- если $k \le 0$ или если *k* больше, чем число точек данных, то функция НАИБОЛЬШИЙ помещает в ячейку значение ошибки #ЧИСЛО!.

Математико-статистическая интерпретация:

Функцию НАИБОЛЬШИЙ удобно применять при выборе значения по его относительному местоположению. Например, ее можно использовать, чтобы определить наилучший, второй, третий и т. д. результат в баллах, показанный при тестировании, измерении и т. п.

Если *n* – число точек в массиве данных, то функция НАИБОЛЬШИЙ(массив;1) находит наибольшее значение, а функция НАИБОЛЬШИЙ(массив;n) – наименьшее.

Например, для исходных данных (*см.* табл. 4.3) в описании функции СРЗНАЧ формула =НАИБОЛЬШИЙ(C40:C56;17) рассчитает значение 206 (Некоузский район), формула =НАИБОЛЬШИЙ(C40:C56; 1) – значение 8872 (Переславский район), а формула =НАИБОЛЬШИЙ(C40:C56;5) – значение 2121 (Некрасовский район).

♦ В примере 4.1 первое по порядку наибольшее значение (показатель *Наибольший (1)* в табл. 4.2) рассчитывается по формуле =НАИБОЛЬШИЙ(B2:B10;1).

Функция НАИМЕНЬШИЙ

См. также НАИБОЛЬШИЙ, МИН.

Синтаксис:

НАИМЕНЬШИЙ (массив; *k*)

Результат:

Находит *k*-е по порядку (начиная с x_{min}) наименьшее значение в множестве данных.

Аргументы:

- *массив:* массив данных, для которых определяется *k*-е наименьшее значение;
- *k:* позиция (начиная с наименьшей) в массиве ячеек данных.

Замечания:

• если массив пуст, то функция НАИМЕНЬШИЙ помещает в ячейку значение ошибки #ЧИСЛО!;

• если $k \leq 0$ или если k больше, чем число точек данных, то функция НАИМЕНЬШИЙ помещает в ячейку значение ошибки #ЧИСЛО!.

Математико-статистическая интерпретация:

Функцию НАИМЕНЬШИЙ удобно применять при выборе значения по его относительному местоположению. Например, ее можно использовать, чтобы определить наихудший, предпоследний и т. д. результат в баллах, показанный при тестировании, измерении и т. п.

Если n – число точек в массиве данных, то функция НАИМЕНЬШИЙ(массив;1) находит наименьшее значение, а функция НАИМЕНЬШИЙ(массив;n) – наибольшее.

Например, для исходных данных (*см.* табл. 4.3) в описании функции СРЗНАЧ формула =НАИМЕНЬШИЙ(C40:C56;1) рассчитает значение 206 (Некоузский район), формула =НАИМЕНЬШИЙ(C40:C56;17) – значение 8872 (Переславский район), а формула =НАИМЕНЬШИЙ(C40:C56;5) значение 632 (Брейтовский район).

♦ В примере 4.1 первое по порядку наименьшее значение (показатель *Наименьший* (1) в табл. 4.2) рассчитывается по формуле =НАИМЕНЬШИЙ(B2:B10;1).

Функция СТЬЮДРАСПОБР

См. описание в подразд. 6.3.8.

♦ В примере 4.1 функция СТЬЮДРАСПОБР используется для нахождения коэффициента доверия t (t-критерия Стьюдента) при расчете предельной ошибки выборки $\Delta_{\bar{x}}$ (показатель *Уровень надежности* в табл. 4.2). Значение коэффициента доверия t рассчитывается по формуле =СТЬЮДРАСПОБР(0,05;B28–1),

где 0,05 – уровень значимости $\alpha = 1 - 0{,}95$ (0,95 – доверительная вероятность, заданная в поле *Уровень надежности* диалогового окна **Описательная статистика**, *см.* рис. 4.2);

B28–1 – число степеней свободы $k = n - 1 = 9 - 1 = 8$ (в ячейке B28 – значение объема выборки n, рассчитываемое по формуле =СЧЕТ(B2:B10)).

4.4. Родственные статистические функции

4.4.1. Функции, родственные функции СРЗНАЧ

Функция СРЗНАЧА

См. также СРЗНАЧ.

Синтаксис:

СРЗНАЧА (значение1, значение2, ...)

Результат:

Вычисляет среднюю арифметическую значений, заданных в списке аргументов, которые могут включать текст и логические значения.

Аргументы:

значение1, значение2, ...: от 1 до 30 аргументов, для которых вычисляется средняя арифметическая.

Помимо чисел в расчете могут участвовать текст и логические значения, такие, как ИСТИНА и ЛОЖЬ.

Замечания:

• аргументы должны быть числами, именами, массивами или ссылками;

• массивы и ссылки, содержащие текст, интерпретируются как 0. Пустой текст («») интерпретируется как 0. Если при расчете не требуется учитывать текстовые значения, следует использовать функцию СРЗНАЧ;

• аргументы, содержащие значение ИСТИНА, интерпретируются как 1. Аргументы, содержащие значение ЛОЖЬ, интерпретируются как 0;

• вычисляя средние значения ячеек, следует учитывать различие между пустыми ячейками и ячейками, содержащими нулевые значения, особенно если не установлен флажок *Нулевые значения* на вкладке ***Вид*** (команда **Параметры...** меню ***Сервис***). Пустые ячейки не учитываются, но нулевые ячейки учитываются.

Математико-статистическая интерпретация:

См. описание функции СРЗНАЧ.

Функция УРЕЗСРЕДНЕЕ

См. также СРЗНАЧ.

Синтаксис:

УРЕЗСРЕДНЕЕ (массив; доля)

Результат:

Рассчитывает среднюю арифметическую усеченного множества данных путем отбрасывания заданного процента данных с экстремальными значениями. В результате из анализа исключается заданный процент выбросов данных.

Аргументы:

• *массив*: массив усредняемых значений;

• *доля*: доля точек данных, исключаемых из вычислений. Например, если доля равна 0,2, то из множества данных, содержащих 40 точек, исключаются 8 точек (40 · 0,2), 4 точки с наибольшими значениями и 4 точки с наименьшими значениями в множестве данных.

Замечания:

• если *доля* < 0 или *доля* > 1, то функция УРЕЗСРЕДНЕЕ помещает в ячейку значение ошибки #ЧИСЛО!;

• функция УРЕЗСРЕДНЕЕ округляет количество отбрасываемых точек данных с недостатком до ближайшего целого, кратного 2. Если доля равна 0,1, то 10 % от 30 точек данных составляют 3 точки, но из соображений симметрии функция УРЕЗСРЕДНЕЕ исключит по одному значению из начала и конца множества.

Математико-статистическая интерпретация:

См. описание функции СРЗНАЧ.

Функция УРЕЗСРЕДНЕЕ рассчитывает значение средней арифметической при уменьшенном размахе вариации (размах вариации рассчитывается как разность между наибольшим и наименьшим значениями признака, *см.* описание функции МАКС).

Размах вариации позволяет судить об устойчивости значений варьирующего признака. Вместе с тем к существенным недостаткам размаха вариации можно отнести то обстоятельство, что очень низкое и очень высокое значения признака по сравнению с основной массой его значений в совокупности могут быть обусловлены какими-либо сугубо случайными обстоятельствами (т. е. эти значения являются аномальными в совокупности или, иначе, *выбросами*). В этих случаях выбросы могут существенно

исказить статистические оценки. Поэтому следует очистить совокупность от выбросов, прежде чем приступать к обработке данных.

Например, для исходных данных (*см.* табл. 4.3) в описании функции СРЗНАЧ при установленной доле 0,2 формула УРЕЗСРЕДНЕЕ(C40:C56;0,2) рассчитает значение 1520 (будут отброшены значения для Некоузского и Переславского районов).

Внимание! Исключать «подозрительные» выбросы из совокупности нужно очень осторожно. Необоснованное исключение «подозрительных» выбросов может привести к существенному искажению статистических оценок. В связи с этим возникает необходимость решения задачи – считать ли данный «подозрительный» выброс принадлежащим данной генеральной совокупности или аномальным, подлежащим исключению из выборки. Подходы к решению этой задачи при нормальном распределении данных можно найти в работе [10].

Исключение крайних оценок (выбросов) широко используется в методах коллективной экспертной оценки, в частности в методе «Дельфи». Так, при рассмотрении оценок группы экспертов оценка, слишком сильно отличающаяся от других, может быть исключена из дальнейшего рассмотрения, несмотря на то, что она может оказаться более верной, чем остальные. Правда, подобные отклонения, по мнению авторов «Дельфи» Т. Гордона и О. Хелмера, компенсируются тем, что эксперта, не согласного с большинством, просят высказать причины несогласия. Все эксперты имеют возможность ознакомиться с его доводами и могут принять во внимание или отвергнуть их, переоценить свое мнение или остаться при нем. Так что оценка, далеко отстоящая от других, отбрасывается фактически лишь в том случае, если эксперту не удается достаточно веско аргументировать свою точку зрения.

Функция СРГАРМ

См. также СРЗНАЧ, СРГЕОМ.

Синтаксис:

СРГАРМ (число1; число2; ...)

Результат:

Рассчитывает среднюю гармоническую множества данных.

Аргументы:

число1, число2, ...: от 1 до 30 аргументов, для которых вычисляется средняя гармоническая.

Замечания:

- аргументы должны быть числами или именами, массивами или ссылками, содержащими числа;
- если аргумент, который является массивом или ссылкой, содержит текстовые, логические значения или пустые ячейки, то такие значения игнорируются, однако ячейки с нулевыми значениями учитываются;
- если какой-либо из аргументов меньше или равен 0, то функция СРГАРМ помещает в ячейку значение ошибки #ЧИСЛО!;
- средняя гармоническая всегда меньше средней геометрической.

Математико-статистическая интерпретация:

Средняя гармоническая – это величина, обратная к средней арифметической обратных величин. Формулы для средней гармонической имеют следующий вид:

$$\overline{h} = \frac{\sum W_i}{\sum \frac{W_i}{x_i}} \quad \text{(взвешенная средняя гармоническая)},$$

где $W_i = x_i f_i$;

$$\overline{h} = \frac{n}{\sum \frac{1}{x_i}} \quad \text{(невзвешенная средняя гармоническая)}.$$

Внимание! Функция СРГАРМ рассчитывает значение *невзвешенной* средней гармонической, ее нельзя использовать для расчета взвешенной средней гармонической.

Формула средней гармонической взвешенной применяется, когда статистическая информация не содержит частот по отдельным вариантам совокупности, а представлена как их произведение.

Рассмотрим расчет взвешенной средней гармонической на примере расчета средней цены на компьютерную энциклопедию «Шедевры русской живописи» (табл. 4.15).

Ячейка D7 содержит формулу {=СУММ (D4:D6)/СУММ(D4::D6/C4:C6)}.

Таблица 4.15

	B	C	D
2	Цена и сумма выручки от продажи CD «Шедевры русской живописи»		
3	Фирма	Цена x_i, руб.	Сумма выручки W_i, руб.
4	«Никита»	290	20300
5	«Кирилл и Мефодий»	270	27000
6	«Рога и копыта»	55	16500
7	Средняя цена (средняя гармоническая)		**135,74**

Примечание. Данная формула вводится как формула массива. Для этого следует активизировать ячейку, в которую необходимо ввести формулу (в нашем случае D7), набрать формулу и нажать комбинацию клавиш **Ctrl+Shift+Enter**. После нажатия указанной комбинации клавиш Microsoft Excel автоматически заключит формулу в фигурные скобки {}.

Внимание! При определении средней цены, используя невзвешенную среднюю арифметическую, получим среднюю, которая не учитывает объема реализации ($f_1 = 70, f_2 = 100, f_3 = 300$), что приводит к неверному результату 203 = (290 + 270 + 55)/3. Неверный результат также даст и формула =СРГАРМ(C4:C6), так как она рассчитывает значение невзвешенной средней гармонической, равное 118,42.

Как видно, средняя гармоническая является превращенной формой средней арифметической. Вместо гармонической всегда можно рассчитать среднюю арифметическую, но для этого сначала нужно определить веса отдельных значений признака f_i.

Средняя гармоническая невзвешенная используется на практике значительно реже. Она применима в том случае, если объемы явлений, т. е. произведения по каждому признаку, равны. Допустим, что у всех трех фирм выручка за реализацию CD одинакова (ячей-

ки D4:D6 в табл. 4.15). Тогда можно применить функцию СРГАРМ, использующую упрощенную формулу

$$\bar{h} = \frac{3}{\frac{1}{290} + \frac{1}{270} + \frac{1}{55}} = 118{,}42.$$

Заметим, что формула {=СУММ(D4:D6)/СУММ(D4:D6/C4::C6)} также рассчитывает значение 118,42 во всех тех случаях, когда значения ячеек D4:D6 равны между собой.

Функция СРГЕОМ

См. также СРЗНАЧ, СРГАРМ.

Синтаксис:

СРГЕОМ (число1; число2; ...)

Результат:

Рассчитывает среднюю геометрическую значений массива положительных чисел.

Аргументы:

число1, число2, ...: от 1 до 30 аргументов, для которых вычисляется средняя геометрическая.

Замечания:

- аргументы должны быть числами или именами, массивами или ссылками, содержащими числа;
- если аргумент, который является массивом или ссылкой, содержит текстовые, логические значения или пустые ячейки, то такие значения игнорируются; однако ячейки с нулевыми значениями учитываются;
- если какой-либо из аргументов меньше или равен 0, то функция СРГЕОМ помещает в ячейку значение ошибки #ЧИСЛО!.

Математико-статистическая интерпретация:

Одной из формул, по которой может рассчитываться средний показатель, является *средняя геометрическая*:

$$\bar{g} = \sqrt[k]{x_1 x_2 x_3 \ldots x_k} = \sqrt[k]{\prod x_i}.$$

Этой средней удобно пользоваться, когда внимание уделяется не абсолютным разностям, а отношениям двух чисел. Поэтому

средняя геометрическая используется в расчетах среднегодовых темпов роста.

Средний темп роста является сводной обобщающей характеристикой интенсивности изменения уровней ряда динамики. Он показывает, во сколько раз в среднем за единицу времени изменился уровень динамического ряда. Необходимость исчисления среднего темпа роста возникает вследствие того, что темпы роста из года в год колеблются. Обычно средний темп роста вычисляется по формуле средней геометрической из цепных коэффициентов роста:

$$\overline{T_3} = \sqrt[n-1]{K_{2/1} K_{3/2} \ldots K_{n/n-1}} = \sqrt[n-1]{\prod K_{i/i-1}}.$$

Рассмотрим использование функции СРГЕОМ на примере расчета среднего темпа роста производства молока в регионе за 1993–1998 гг. (табл. 4.16).

Таблица 4.16

	B	C	D
2	Динамика производства молока в регионе за 1993–1995 гг. (цифры условные)		
3			
4	Год	Произведено молока, тыс. т	Коэффициент роста $K_{i/i-1}$
5	1994	310,12	–
6	1995	321,50	1,04
7	1996	340,70	1,06
8	1997	315,40	0,93
9	1998	335,90	1,06
10	Средний темп роста производства молока за 1993–1995 гг.		**1,02**

В среднем за год производство молока в регионе за период с 1993 г. по 1998 г. увеличилось в 1,02 раза.

Ячейки D6:D9 содержат формулы цепного коэффициента роста (например, в ячейке D6 содержится формула =C6/C5). Ячейка D10 содержит формулу =СРГЕОМ(D6:D9).

4.4.2. Функции, родственные функции МЕДИАНА

Функция КВАРТИЛЬ

См. также МЕДИАНА, ПЕРСЕНТИЛЬ.

Синтаксис:

КВАРТИЛЬ (массив; часть)

Результат:

Рассчитывает квартиль для множества данных.

Аргументы:

• *массив*: массив ячеек с числовыми значениями, для которых определяются значения квартилей;

• *часть*: аргумент, определяющий, что будет рассчитывать функция КВАРТИЛЬ.

Замечания:

• функция КВАРТИЛЬ рассчитывает:

♦ минимальное значение, если аргумент *часть* = 0;

♦ первую квартиль (25-ю персентиль), если аргумент *часть* = 1;

♦ значение медианы (50-ю персентиль), если аргумент *часть* = 2;

♦ третью квартиль (75-ю персентиль), если аргумент *часть* = 3;

♦ максимальное значение, если аргумент *часть* = 4;

• если массив пуст или содержит более 8191 точки данных, то функция КВАРТИЛЬ помещает в ячейку значение ошибки #ЧИСЛО!;

• если аргумент *часть* не целое число, то он усекается;

• если аргумент *часть* < 0 или *часть* > 4, то функция КВАРТИЛЬ помещает в ячейку значение ошибки #ЧИСЛО!;

• функции МИН, МЕДИАНА и МАКС рассчитывают то же самое значение, что и функция КВАРТИЛЬ, если аргумент *часть* равен 0, 2 или 4 соответственно.

Математико-статистическая интерпретация:

Квартили представляют собой значения признака, делящие ранжированную совокупность на четыре равновеликие части. Различают квартиль нижний (Q_1), отделяющий 1/4 часть совокупности с наименьшими значениями признака, и квартиль верхний (Q_3), отделяющий 1/4 часть с наибольшими значениями признака. Средним квартилем (Q_2) является медиана.

Квартиль часто используется при анализе продаж, чтобы разбить генеральную совокупность на группы. Например, можно использовать функцию КВАРТИЛЬ, чтобы найти 25% наиболее доходных предприятий.

Приведем результаты, рассчитанные функцией КВАРТИЛЬ на основании исходных данных из табл. 4.7 (*см.* описание функции МОДА). Функция КВАРТИЛЬ не требует предварительной ранжировки данных, она проводит ее автоматически.

Формула	Результат
=КВАРТИЛЬ(C4:C8;0)	48,00
=КВАРТИЛЬ(C4:C8;1)	57,00
=КВАРТИЛЬ(C4:C8;2)	60,00
=КВАРТИЛЬ(C4:C8;3)	77,00
=КВАРТИЛЬ(C4:C8;4)	95,00

В отличие от дискретных вариационных рядов определение квартилей по *интервальным* рядам требует проведения расчетов по следующим формулам:

$$Q_1 = x_{Q_1} + i\frac{0,25\sum f_i - S_{Q_1-1}}{f_{Q_1}};$$

$$Q_3 = x_{Q_3} + i\frac{0,75\sum f_i - S_{Q_3-1}}{f_{Q_3}},$$

где x_{Q_1} – нижняя граница интервала, содержащего нижний квартиль (интервал определяется по накопленной частоте, первой превышающей 25 %);

x_{Q_3} – нижняя граница интервала, содержащего верхний квартиль (интервал определяется по накопленной частоте, первой превышающей 75 %);

i – величина интервала;

f_{Q_1}, f_{Q_3} – частота интервала, содержащего нижний квартиль и верхний квартиль соответственно;

$S_{Q_{1-1}}, S_{Q_{3-1}}$ – накопленная частота интервала, предшествующего интервалу, содержащему нижний квартиль, верхний квартиль соответственно.

В табл. 4.17 интервалом, содержащим нижний квартиль, является интервал 4–6 лет, а первый квартиль имеет значение 5,58 года.

Содержимое ячеек в табл. 4.17 и 4.6 (*см.* описание функции МЕДИАНА) аналогично, за исключением следующих ячеек:

Таблица 4.17

	B	C	D
2	Научный стаж сотрудников НИЦ, лет	Число сотрудников, f_i	Накопленная частота, S_i
3	До 4	14	14
4	4–6	33	47
5	6–8	30	77
6	8–10	45	122
7	10–12	21	143
8	Свыше 12	17	160
9	Итого	160	
10	25 % числа сотрудников	40	
11	Смещение на max ≤N/4	1	
12	Значение max≤N/4	14	
13	Смещение на первый квартильный интервал	2	
14	Накопленная частота первого квартильного интервала	47	
15	Первый квартильный интервал	4–6	
16	Нижняя граница первого квартильного интервала	4	
17	Значение накопленной частоты предшествующего интервала	14	
18	Первый квартиль	**5,58**	

• ячейка C10 содержит формулу =C9/4 – рассчитывается четвертая часть численности совокупности (25 % числа сотрудников);

• ячейка C18 содержит формулу =C16+2*((C9/4–C17)/C14) – вычисляется значение первого квартиля.

Второй квартиль совпадает с медианой (*см.* описание функции МЕДИАНА) и равен 8,13. Верхний (третий) квартиль содержится в интервале 8–10 лет и равен 9,91.

Функция ПЕРСЕНТИЛЬ

См. также МЕДИАНА, КВАРТИЛЬ.

Синтаксис:

ПЕРСЕНТИЛЬ (массив; *k*)

Результат:

Рассчитывает *k*-ю перцентиль для множества данных.

Аргументы:

• *массив*: массив ячеек с числовыми значениями, для которых определяются значения перцентилей;

• *k*: значение перцентили в интервале от 0 до 1 включительно.

Замечания:

• если массив пуст или содержит более 8191 точки данных, то функция ПЕРСЕНТИЛЬ помещает в ячейку значение ошибки #ЧИСЛО!;

• если *k* не является числом, то функция ПЕРСЕНТИЛЬ помещает в ячейку значение ошибки #ЗНАЧ!;

• если $k < 0$ или $k > 1$, то функция ПЕРСЕНТИЛЬ помещает в ячейку значение ошибки #ЧИСЛО!;

• если *k* не кратно $1/(n-1)$, то функция ПЕРСЕНТИЛЬ производит интерполяцию для определения значения *k*-й перцентили.

Математико-статистическая интерпретация:

Кроме квартилей в вариационных рядах распределения могут определяться децили и перцентили. Последние также иногда называют персентилями или процентилями. Децили делят ранжированную совокупность на десять равновеликих частей, а перцентили – на сто. Соотношения медианы, квартилей, децилей и перцентилей представлены на рис. 4.5.

Перцентили применяются лишь при необходимости подробного изучения структуры вариационного ряда.

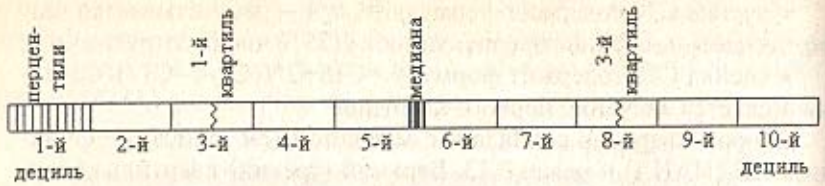

Рис. 4.5

25-я перцентиль является 1-м (нижним) квартилем, 50-я – 2-м квартилем (медианой), 75-я – 3-м (верхним) квартилем.

Приведем некоторые результаты, рассчитанные функцией ПЕРСЕНТИЛЬ на основании исходных данных из табл. 4.7 (*см.* описание функции МОДА). Функция ПЕРСЕНТИЛЬ не требует предварительной ранжировки данных, она проводит ее автоматически.

Формула	Результат
=ПЕРСЕНТИЛЬ(C4:C8; 0)	48,00
=ПЕРСЕНТИЛЬ(C4:C8;0,25)	57,00
=ПЕРСЕНТИЛЬ(C4:C8;0,50)	60,00
=ПЕРСЕНТИЛЬ(C4:C8;0,75)	77,00
=ПЕРСЕНТИЛЬ(C4:C8;1)	95,00

В отличие от дискретных вариационных рядов вычисление перцентилей по *интервальным* рядам требует проведения определенных расчетов. Вычисляются они по той же схеме, что и медиана, и квартили:

$$p_1 = x_{p_1} + i\frac{0,01\sum f_i - S_{p_1-1}}{f_{p_1}};$$

$$p_2 = x_{p_2} + i\frac{0,02\sum f_i - S_{p_2-1}}{f_{p_2}} \text{ и т. д.,}$$

где x_{p_i} – нижняя граница интервала, содержащего i-ю перцентиль;

i – величина интервала;

f_{p_i} – частота интервала, содержащего i-ю перцентиль;

S_{p_i-1} – накопленная частота интервала, предшествующего интервалу, содержащему i-ю перцентиль.

Для данных табл. 4.17 (*см.* описание функции КВАРТИЛЬ) 25-я перцентиль равна 5,58, 40-я перцентиль – 7,13, 50-я перцентиль – 8,13, 60-я перцентиль – 8,84, 75-я перцентиль – 9,91.

4.4.3. Функции, родственные функциям ДИСП и СТАНДОТКЛОН

Функция ДИСПА

См. также ДИСП, ДИСПРА, КВАДРОТКЛ, СРОТКЛ, СТАНДОТКЛОНА.

Синтаксис:

ДИСПА (значение1; значение2; ...)

Результат:

Оценивает генеральную дисперсию *по выборке*, заданной аргументами, которые могут включать текстовые и логические значения.

Аргументы:

значение1, значение2, ...: от 1 до 30 аргументов, соответствующих выборке из генеральной совокупности.

В расчете помимо численных значений учитываются также текстовые и логические значения, такие, как ИСТИНА или ЛОЖЬ.

Замечания:

- функция ДИСПА предполагает, что аргументы являются только выборкой из генеральной совокупности. Если данные представляют всю генеральную совокупность, нужно вычислять дисперсию, используя функцию ДИСПРА;
- аргументы, содержащие значение ИСТИНА, интерпретируются как 1; аргументы, содержащие текст или значение ЛОЖЬ, интерпретируются как 0. Если текстовые и логические значения должны игнорироваться, следует использовать функцию ДИСП.

Математико-статистическая интерпретация:

См. описание функции ДИСП.

Функция ДИСПР

См. также ДИСП, ДИСПРА, КВАДРОТКЛ, СРОТКЛ, СТАНДОТКЛОНП.

Синтаксис:

ДИСПР (число1; число2; ...)

Результат:

Вычисляет дисперсию по *генеральной* совокупности.

Аргументы:

число1, число2, ...: от 1 до 30 аргументов, соответствующих генеральной совокупности.

Замечания:

- логические значения, например ИСТИНА и ЛОЖЬ, а также текст игнорируются. Если они не должны игнорироваться, следует использовать функцию ДИСПРА;
- функция ДИСПР предполагает, что аргументы представляют всю генеральную совокупность. Если данные представляют только выборку из генеральной совокупности, то дисперсию следует вычислять, используя функцию ДИСП.

Математико-статистическая интерпретация:

Дисперсия (от лат. dispersio – рассеяние) – числовая характеристика случайной величины, характеризующая рассеяние ее возможных значений около математического ожидания. В теории вероятностей дисперсия вычисляется через центральный момент 2-го порядка:

$$\sigma^2[X] = D[X] = \mu_2[X] = M[(X - m_x)^2].$$

Аналогично этому статистическая дисперсия определяется через статистический (эмпирический) центральный момент 2-го порядка, представляет собой средний квадрат отклонений индивидуальных значений признака от их средней величины и вычисляется в зависимости от исходных данных по формулам *невзвешенной* (*простой*) и *взвешенной* дисперсий:

$$\sigma^2 = \frac{\sum (x_i - \overline{x})^2}{n} \text{ (простая дисперсия);}$$

$$\sigma^2 = \frac{\sum (x_i - \overline{x})^2 f_i}{\sum f_i} \text{ (взвешенная дисперсия).}$$

Внимание! Функция ДИСПР рассчитывает дисперсию при условии, что исходные данные образуют генеральную совокупность. В случае если совокупность является выборочной, необходимо воспользоваться функцией ДИСП.

Используем исходные данные из табл. 4.10 (*см.* описание функции СТАНДОТКЛОН), предполагая, что они образуют генеральную совокупность. Тогда невзвешенная дисперсия будет определяться формулой =ДИСПР(C4:C9) и равняться 7466,67 (сравните со значением 8960, рассчитываемым функцией ДИСП).

Взвешенная дисперсия находится по аналогии с расчетом взвешенной средней арифметической (*см.* описание функции СРЗНАЧ).

Функция ДИСПРА

См. также ДИСПР, ДИСПА, КВАДРОТКЛ, СРОТКЛ, СТАНДОТКЛОНПА.

Синтаксис:

ДИСПРА (значение1; значение2; ...)

Результат:

Вычисляет дисперсию по генеральной совокупности, заданной аргументами, которые могут включать текстовые и логические значения.

Аргументы:

значение1, значение2, ...: от 1 до 30 аргументов, соответствующих генеральной совокупности.

В расчете помимо численных значений учитываются также текстовые и логические значения, такие, как ИСТИНА или ЛОЖЬ.

Замечания:

- функция ДИСПРА предполагает, что аргументы представляют всю генеральную совокупность. Если данные представляют только выборку из генеральной совокупности, то дисперсию следует вычислять, используя функцию ДИСПА;
- аргументы, содержащие значение ИСТИНА, интерпретируются как 1, аргументы, содержащие текст или значение ЛОЖЬ, интерпретируются как 0. Если текстовые и логические значения должны игнорироваться, следует использовать функцию ДИСПР.

Математико-статистическая интерпретация:

См. описание функции ДИСПР.

Функция СТАНДОТКЛОНА

См. также ДИСПА, КВАДРОТКЛ, СРОТКЛ, СТАНДОТКЛОН, СТАНДОТКЛОНПА.

Синтаксис:

СТАНДОТКЛОНА (значение1; значение2; ...)

Результат:

Оценивает генеральное стандартное отклонение по выборке, заданной аргументами, которые могут включать текстовые и логические значения.

Аргументы:

значение1, значение2, ...: от 1 до 30 аргументов, соответствующих выборке из генеральной совокупности.

В расчете помимо численных значений учитываются также текстовые и логические значения, такие, как ИСТИНА или ЛОЖЬ.

Замечания:

• функция СТАНДОТКЛОНА предполагает, что аргументы являются только выборкой из генеральной совокупности. Если данные представляют всю генеральную совокупность, то стандартное отклонение следует вычислять с помощью функции СТАНДОТКЛОНПА;

• аргументы, содержащие значение ИСТИНА, интерпретируются как 1. Аргументы, содержащие значение ЛОЖЬ, интерпретируются как 0. Если текстовые и логические значения должны игнорироваться, следует использовать функцию СТАНДОТКЛОН.

Математико-статистическая интерпретация:

См. описание функций ДИСП, СТАНДОТКЛОН и СТАНДОТКЛОНП.

Функция СТАНДОТКЛОНП

См. также ДИСПР, КВАДРОТКЛ, СРОТКЛ, СТАНДОТКЛОН, СТАНДОТКЛОНПА.

Синтаксис:

СТАНДОТКЛОНП (число1; число2; ...)

Результат:

Вычисляет стандартное отклонение по *генеральной* совокупности.

Аргументы:

число1, число2, ...: от 1 до 30 аргументов, соответствующих генеральной совокупности.

Замечания:

• функция СТАНДОТКЛОНП предполагает, что аргументы образуют всю генеральную совокупность. Если данные являются только выборкой из генеральной совокупности, то стандартное отклонение следует вычислять с использованием функции СТАНДОТКЛОН;

• для больших выборок функции СТАНДОТКЛОН и СТАНДОТКЛОНП рассчитывают примерно равные значения;

• логические значения, такие, как ИСТИНА или ЛОЖЬ, а также текст игнорируются. Если текстовые и логические значения игнорироваться не должны, следует использовать функцию рабочего листа СТАНДОТКЛОНПА.

Математико-статистическая интерпретация:

См. описание функции ДИСПР.

Дисперсия имеет размерность квадрата вариант. Для наглядной характеристики меры вариации удобнее пользоваться величиной, размерность которой совпадает с размерностью вариант. Для этого из дисперсии извлекают квадратный корень. Полученная величина называется *стандартным отклонением* σ (иначе, *среднеквадратичным отклонением*). Оно выражается в тех же единицах измерения, что и признак (тоннах, рублях, метрах, процентах и т. д.).

Формулы для стандартного отклонения имеют следующий вид:

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}}$$ (простое стандартное отклонение);

$$\sigma = \sqrt{\frac{\sum (x_i - \bar{x})^2 f_i}{\sum f_i}}$$ (взвешенное стандартное отклонение).

Внимание! Функция СТАНДОТКЛОНП рассчитывает стандартное отклонение при условии, что исходные данные образуют *генеральную* совокупность. В случае если совокупность является *выборочной*, необходимо использовать функцию СТАНДОТКЛОН.

Используем исходные данные из табл. 4.10 (*см.* описание функции СТАНДОТКЛОН), предполагая, что они образуют гене-

ральную совокупность. Тогда стандартное отклонение будет определяться формулой =СТАНДОТКЛОНП(С4:С9) и равняться 86,41 (сравните со значением 94,66, рассчитываемым функцией СТАНДОТКЛОН).

Функция СТАНДОТКЛОНПА

См. также ДИСПРА, КВАДРОТКЛ, СРОТКЛ, СТАНДОТКЛОНА, СТАНДОТКЛОНП.

Результат:

Вычисляет стандартное отклонение по генеральной совокупности, заданной аргументами, которые могут включать текстовые и логические значения.

Синтаксис:

СТАНДОТКЛОНПА (значение1; значение2; ...)

Аргументы:

значение1, значение2, ...: от 1 до 30 аргументов, соответствующих генеральной совокупности.

В расчете помимо численных значений учитываются также текстовые и логические значения, такие, как ИСТИНА или ЛОЖЬ.

Замечания:

- функция СТАНДОТКЛОНПА предполагает, что аргументы образуют всю генеральную совокупность. Если данные являются только выборкой из генеральной совокупности, то стандартное отклонение следует вычислять с использованием функции СТАНДОТКЛОНА;
- аргументы, содержащие значение ИСТИНА, интерпретируются как 1, аргументы, содержащие значение ЛОЖЬ, интерпретируются как 0. Если текстовые и логические значения должны игнорироваться, следует использовать функцию СТАНДОТКЛОНП;
- для больших выборок функции СТАНДОТКЛОНА и СТАНДОТКЛОНПА рассчитывают примерно равные значения.

Математико-статистическая интерпретация:

См. описание функции СТАНДОТКЛОНП.

Функция КВАДРОТКЛ

См. также ДИСПР, СТАНДОТКЛОН.

Синтаксис:

КВАДРОТКЛ (число1; число2; ...)

Результат:

Рассчитывает сумму квадратов отклонений точек данных от их средней арифметической.

Аргументы:

число1, число2, ...: от 1 до 30 аргументов, для которых вычисляется сумма квадратов отклонений.

Замечания:

- аргументы должны быть числами или именами, массивами или ссылками, содержащими числа;
- если аргумент, который является массивом или ссылкой, содержит текстовые, логические значения или пустые ячейки, то такие значения игнорируются; однако ячейки с нулевыми значениями учитываются.

Математико-статистическая интерпретация:

Формула для квадратичного отклонения имеет следующий вид:

$$\theta = \sum (x_i - \bar{x})^2.$$

Как самостоятельная мера вариации квадратичное отклонение в экономической статистике используется редко. Оно входит составной частью в выражения дисперсии и стандартного отклонения (*см.* описание функций ДИСПР, СТАНДОТКЛОНП). В качестве самостоятельной меры вариации квадратичное отклонение применяется в некоторых разделах статистической физики, в частности при оценке флуктуаций хаотического теплового движения частиц.

Используя данные, приведенные в табл. 4.10 (*см.* описание функции СТАНДОТКЛОН), по формуле =КВАДРОТКЛ(С4:С9) получим квадратичное отклонение 44800.

Внимание! Не путать квадратичное отклонение со среднеквадратичным (стандартным) отклонением, вычисляемым функцией СТАНДОТКЛОН.

Функция СРОТКЛ

См. также ДИСП, КВАДРОТКЛ, СТАНДОТКЛОН.

Синтаксис:

СРОТКЛ (число1; число2; ...)

Результат:

Вычисляет среднее линейное отклонение в множестве данных.

Аргументы:

число1, число2, ...: от 1 до 30 аргументов, для которых определяется среднее абсолютных отклонений.

Замечания:

• аргументы должны быть числами или именами, массивами или ссылками, содержащими числа;

• если аргумент, который является массивом или ссылкой, содержит текстовые, логические значения или пустые ячейки, то такие значения игнорируются; однако ячейки, которые содержат нулевые значения, учитываются.

Математико-статистическая интерпретация:

Среднее линейное отклонение определяется как средняя арифметическая из абсолютных значений отклонений вариант x_i от $\bar{x}$. В зависимости от характера имеющихся данных среднее линейное отклонение может быть *невзвешенным* и *взвешенным* (аналогично средней арифметической).

Функция СРОТКЛ рассчитывает значение *невзвешенного* среднего линейного отклонения по формуле

$$\bar{d} = \frac{\sum |x_i - \bar{x}|}{n}.$$

Для исходных данных из табл. 4.3 (*см.* описание функции СРЗНАЧ) среднее линейное отклонение объема индивидуального жилищного строительства составит 1505,2. Ячейка С57 в этом случае будет содержать формулу =СРОТКЛ(С40:С56).

Взвешенное среднее линейное отклонение можно найти аналогично взвешенной средней арифметической (*см.* описание функции СРЗНАЧ), а его значение определяется формулой

$$\bar{d} = \frac{\sum |x_i - \bar{x}| f_i}{\sum f_i}.$$

Таким образом, среднее линейное отклонение дает обобщенную характеристику степени колеблемости признака в совокупности. Однако этот показатель в статистической практике применяют редко, так как во многих случаях он не устанавливает степень рассеивания. На практике меру вариации более объективно

отражает показатель дисперсии (*см.* описание функций ДИСП, ДИСПР, СТАНДОТКЛОН, СТАНДОТКЛОНП).

4.4.4. Функции, родственные функции СЧЕТ

Функция СЧЕТЗ

См. также СЧЕТ.

Синтаксис:

СЧЕТЗ (значение1; значение2; ...)

Результат:

Рассчитывает количество непустых значений в списке аргументов.

Аргументы:

значение1, значение2, ...: от 1 до 30 аргументов, количество которых требуется сосчитать.

Замечания:

• в функции СЧЕТЗ в отличие от функции СЧЕТ значением считается значение любого типа, включая пустую строку («»), но не пустые ячейки;

• если аргументом является массив, то пустые ячейки в массиве игнорируются.

Математико-статистическая интерпретация:

Функция СЧЕТЗ используется для подсчета в массиве количества ячеек, содержащих данные.

Для исходных данных (*см.* табл. 4.13) в описании функции СКОС формулы =СЧЕТ(D3:D11) и =СЧЕТ(C3:C11) рассчитывают значение 9 (сравните с функцией СЧЕТ).

4.4.5. Функции, родственные функции МИН

Функция МИНА

См. также МИН, МАКС, МАКСА.

Синтаксис:

МИНА (значение1; значение2; ...)

Результат:

Находит наименьшее значение в множестве данных, которые могут включать текстовые и логические значения.

Аргументы:

значение1, значение2, ...: от 1 до 30 аргументов, среди которых ищется минимальный.

Замечания:

- можно задавать аргументы, которые являются числами, пустыми ячейками, логическими значениями или текстовыми представлениями чисел. Если логические значения и тексты должны игнорироваться, то следует использовать функцию МИН;
- если аргументом является массив или ссылка, учитываются только значения массива или ссылки. Пустые ячейки и тексты в массиве или ссылке игнорируются;
- аргументы, содержащие значение ИСТИНА, интерпретируются как 1; аргументы, содержащие текст или значение ЛОЖЬ, интерпретируются как 0;
- если аргументы не содержат значений, то функция МИНА помещает в ячейку значение 0.

Математико-статистическая интерпретация:

См. описание функции МИН.

4.4.6. Функции, родственные функции МАКС

Функция МАКСА

См. также МАКС, МИН, МИНА.

Синтаксис:

МАКСА (*значение*1; *значение*2; ...)

Результат:

Находит наибольшее значение в множестве данных, которые могут включать текстовые и логические значения.

Аргументы:

значение1, значение2, ...: от 1 до 30 аргументов, среди которых ищется максимальный.

Замечания:

- можно задавать аргументы, которые являются числами, пустыми ячейками, логическими значениями или текстовыми пред-

ставлениями чисел. Если логические значения и тексты должны игнорироваться, то следует использовать функцию МАКС;
- если аргументом является массив или ссылка, учитываются только значения массива или ссылки. Пустые ячейки и тексты в массиве или ссылке игнорируются;
- аргументы, содержащие значение ИСТИНА, интерпретируются как 1; аргументы, содержащие текст или значение ЛОЖЬ, интерпретируются как 0;
- если аргументы не содержат значений, то функция МАКСА помещает в ячейку значение 0.

Математико-статистическая интерпретация:

См. описание функции МАКС.

ГЛАВА 5 Ранг и персентиль

5.1. Краткие сведения из теории статистики

При проведении анализа взаимного расположения значений признака в наборе данных наряду с такими понятиями, как мода, медиана, квартиль, квинтель, дециль, персентиль (*см.* главу 4), пользуются также понятиями *ранга* и *процентранга*.

Под *рангом* (R) понимают номер (порядковое место) значения случайной величины в наборе данных. Правила присвоения рангов состоят в следующем:

1) если в наборе данных все числа разные, то каждому числу x_i присваивается уникальный ранг R_i;

2) если в наборе данных встречается группа из k одинаковых чисел $x_i = x_{i+1} = x_{i+2} = ... = x_{i+k}$, то ранг у них одинаковый и равен рангу первого числа из этой группы R_i. Число, следующее за этой группой, получает ранг, равный R_{i+k};

3) если данные упорядочены *в порядке убывания*, то:

а) максимальное значение в наборе данных имеет ранг, равный 1;

б) минимальное значение в наборе данных имеет наибольшее значение ранга, равное $n - k_{min} + 1$, где n – количество данных в наборе, k_{min} – количество повторяющихся минимальных значений в наборе данных;

4) если данные упорядочены *в порядке возрастания*, то:

а) минимальное значение в наборе данных имеет ранг, равный 1;

б) максимальное значение в наборе данных имеет наибольшее значение ранга, равное $n - k_{max} + 1$, где n – количество данных в наборе, k_{max} – количество повторяющихся максимальных значений в наборе данных.

Под *процентрангом* (T) понимают процентное отношение для каждого значения в наборе данных. Правила вычисления процентранга состоят в следующем:

1) для чисел, значения которых *не повторяются* в наборе данных, применяется формула

$$T_i = \frac{(n - R_i)}{n-1} \cdot 100\%,$$

где n – количество данных в наборе;

R_i – ранг i-го числа, рассчитанный при условии упорядочения данных *по убыванию*.

2) для чисел, значения которых *повторяются* в наборе данных, применяется формула

$$T_i = \frac{(n - R_i - (k_i - 1))}{n-1} \cdot 100\%,$$

где n – количество данных в наборе;

R_i – ранг i-го числа, рассчитанный при условии упорядочения данных *по убыванию*;

k_i – количество повторяющихся значений i-го числа в наборе данных.

Ранги характеризуют взаимное расположение значений признака в наборе данных, а также находят практическое применение в *непараметрических* методах оценки взаимосвязи социально-экономических явлений и процессов (*см.* главы 13 и 14).

В частности, ранги входят в формулу расчета коэффициента Спирмена* (ρ):

---

* (Spearman) Спирмен Чарльз Эдуард (1863–1945) – английский психолог, разработал основы факторного анализа в психологии.

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

где d_i^2 – квадраты разности рангов взаимосвязанных величин X и Y;

n – число наблюдений (число пар рангов).

Коэффициент Спирмена может быть использован для определения тесноты связи как между количественными, так и между качественными признаками при условии, что значения этих признаков могут быть упорядочены по убыванию или возрастанию.

5.2. Справочная информация по технологии работы

Режим работы «Ранг и персентиль» служит для генерации таблицы, содержащей порядковые и процентные ранги для каждого значения из набора данных, при этом данные упорядочиваются в *порядке убывания*.

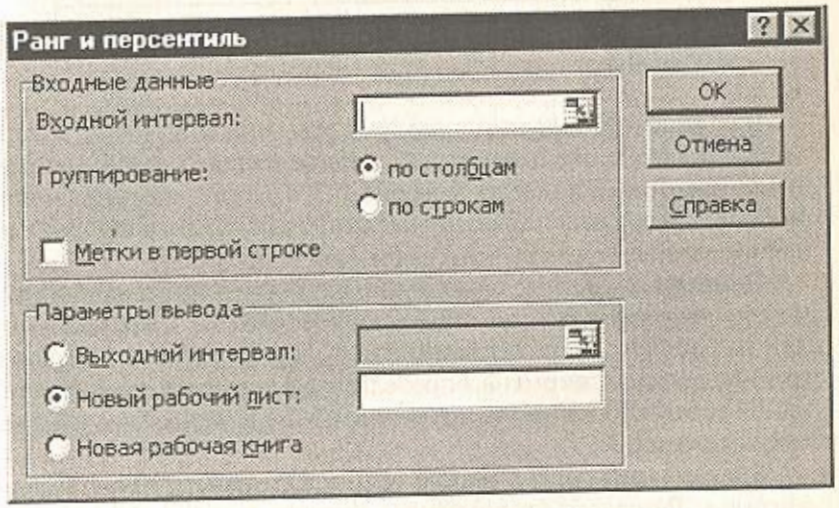

Рис. 5.1

В диалоговом окне данного режима (рис. 5.1) задаются следующие параметры (*см.* подразд. 1.1.2):

1. *Входной интервал.*
2. *Группирование.*
3. *Метки в первой строке/Метки в первом столбце.*
4. *Выходной интервал/Новый рабочий лист/Новая рабочая книга.*

Пример 5.1. Данные о количестве проданных спортивных костюмов «Reebok» фирмой «Чемпион» за 2000 г. приведены в табл. 5.1, сформированной на рабочем листе Microsoft Excel. Необходимо провести количественный анализ относительного взаиморасположения данных в представленном наборе.

Таблица 5.1

	B	C
2	Спрос на спортивные костюмы «Reebok» в фирме «Чемпион» (за 2000 г.)	
3	Размер костюма	Число купленных костюмов
4	46	57
5	48	48
6	50	95
7	52	60
8	54	77

Для решения задачи используем режим работы «Ранг и персентиль». Значения параметров, установленных в одноименном диалоговом окне, приведены на рис. 5.2, а рассчитанные в данном режиме порядковые и процентные ранги для каждого значения из набора данных – в табл. 5.2 (графы *Ранг* и *Процент*).

Пример 5.2. Данные о предприятиях города, выставивших акции на чековый аукцион, приведены в табл. 5.3, сформированной на рабочем листе Microsoft Excel [12]. Требуется с помощью коэффициента Спирмена определить зависимость между величиной уставного капитала предприятий X и количеством выставленных акций Y.

Для решения задачи используем режим работы «Ранг и персентиль». Результаты выполнения данного режима приведены в табл. 5.4.

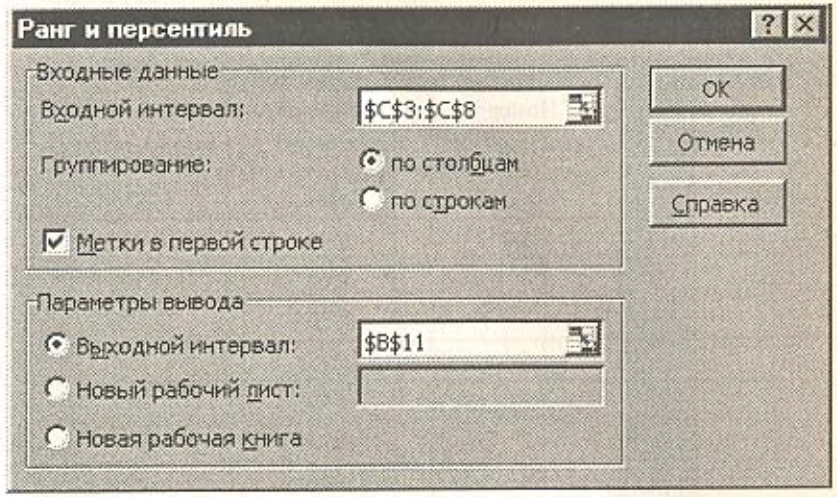

Рис. 5.2

Таблица 5.2

	B	C	D	E
11	*Точка*	*Число купленных костюмов*	*Ранг*	*Процент*
12	3	95	1	100,00%
13	5	77	2	75,00%
14	4	60	3	50,00%
15	1	57	4	25,00%
16	2	48	5	,00%

По данным сгенерированной табл. 5.4 заполняем в табл. 5.5 графы *Ранг* R_X и *Ранг* R_Y, на основании которых производим вычисления квадратов разности рангов d_i^2.

Заключительным этапом решения задачи является вычисление коэффициента Спирмена по формуле

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

Таблица 5.3

	B	C	D
21	Номер предприятия	Уставный капитал, млн руб. X	Число выставленных акций Y
22	1	2954	856
23	2	1605	930
24	3	4102	1563
25	4	2350	682
26	5	2625	616
27	6	1795	495
28	7	2813	815
29	8	1751	858
30	9	1700	467
31	10	2264	661

Таблица 5.4

	B	C	D	E	F	G	H	I
35	*Точка*	*Столбец 1*	*Ранг*	*Процент*	*Точка*	*Столбец 2*	*Ранг*	*Процент*
36	3	4102	1	100,00%	3	1563	1	100,00%
37	1	2954	2	88,80%	2	930	2	88,80%
38	7	2813	3	77,70%	8	858	3	77,70%
39	5	2625	4	66,60%	1	856	4	66,60%
40	4	2350	5	55,50%	7	815	5	55,50%
41	10	2264	6	44,40%	4	682	6	44,40%
42	6	1795	7	33,30%	10	661	7	33,30%
43	8	1751	8	22,20%	5	616	8	22,20%
44	9	1700	9	11,10%	6	495	9	11,10%
45	2	1605	10	,00%	9	467	10	,00%

подставляя в которую исходные и рассчитанные данные задачи, получим

Таблица 5.5

	B	C	D	E	F	G
21	Номер предприятия	Уставный капитал X, млн руб.	Число выставленных акций Y	Ранг R_X	Ранг R_Y	Квадрат разности рангов $d_i^2 = (R_X - R_Y)^2$
22	1	2954	856	2	4	4
23	2	1605	930	10	2	64
24	3	4102	1563	1	1	0
25	4	2350	682	5	6	1
26	5	2625	616	4	8	16
27	6	1795	495	7	9	4
28	7	2813	815	3	5	4
29	8	1751	858	8	3	25
30	9	1700	467	9	10	1
31	10	2264	661	6	7	1
32					Σ =	120

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 120}{10(10^2-1)} = 0{,}27.$$

Значение коэффициента Спирмена $\rho = 0{,}27$ свидетельствует о слабой связи между рассматриваемыми признаками.

5.3. Статистические функции, связанные с режимом «Ранг и персентиль»

Функция РАНГ

Синтаксис:
РАНГ (число; ссылка; порядок)

Результат:

Рассчитывает порядковый ранг числа в наборе данных.

Аргументы:

• *число*: число, для которого определяется ранг;

• *ссылка*: массив исходных данных (нечисловые значения в массиве игнорируются);

• *порядок*: число, определяющее способ упорядочения.

Замечания:

• если аргумент *порядок* = 0 или опущен, то Microsoft Excel определяет ранг числа, упорядочивая исходный набор данных в порядке убывания;

• если аргумент *порядок* является любым ненулевым числом, то Microsoft Excel определяет ранг числа, упорядочивая исходный набор данных в порядке возрастания.

Математико-статистическая интерпретация:

Ранг числа – это его порядковый номер относительно других чисел в наборе данных.

Правила присвоения ранга рассмотрены в подразд. 5.1.

Функция РАНГ не требует предварительной ранжировки данных, она проводит ее автоматически: при аргументе *порядок* = 0 – *в порядке убывания*, при аргументе *порядок* ≠ 0 – *в порядке возрастания*.

Представим результаты, рассчитанные функцией РАНГ на основании исходных данных из табл. 5.1 при аргументе *порядок* = 0.

Формула	Результат
=РАНГ(57;C4:C8;0)	4
=РАНГ(48;C4:C8;0)	5
=РАНГ(95;C4:C8;0)	1
=РАНГ(60;C4:C8;0)	3
=РАНГ(77;C4:C8;0)	2

Представим результаты, рассчитанные функцией РАНГ на основании исходных данных из табл. 5.1 при аргументе *порядок* = 1.

Примечание. О статистической функции ПЕРСЕНТИЛЬ *см.* в подразд. 4.4.2.

Формула	Результат
=РАНГ(57;C4:C8;1)	2
=РАНГ(48;C4:C8;1)	1
=РАНГ(95;C4:C8;1)	5
=РАНГ(60;C4:C8;1)	3
=РАНГ(77;C4:C8;1)	4

Функция ПРОЦЕНТРАНГ

Синтаксис:

ПРОЦЕНТРАНГ (массив; x; разрядность)

Результат:

Рассчитывает процентный ранг числа в наборе данных.

Аргументы:

• *массив*: массив исходных данных;

• *x*: значение, для которого вычисляется процентное содержание;

• *разрядность*: необязательное значение, которое определяет количество значащих цифр в рассчитываемой величине процентного содержания значения. Если этот аргумент опущен, то функция ПРОЦЕНТРАНГ использует три цифры (0, ###).

Замечания:

• если массив пуст, то функция ПРОЦЕНТРАНГ помещает в ячейку значение ошибки #ЧИСЛО!;

• если аргумент *разрядность* < 1, то функция ПРОЦЕНТРАНГ помещает в ячейку значение ошибки #ЧИСЛО!;

• если аргумент *x* не соответствует ни одному из значений аргумента *массив*, то функция ПРОЦЕНТРАНГ производит интерполяцию и рассчитывает корректное значение процентного содержания.

Математико-статистическая интерпретация:

Функция ПРОЦЕНТРАНГ используется для оценки относительного положения точки данных в множестве данных, например, чтобы оценить положение подходящего результата среди всех результатов тестирования.

Для точек, совпадающих с точками множества данных, функция ПРОЦЕНТРАНГ производит расчет по правилам, рассмотренным в подразд. 5.1.

Для точек, не совпадающих с точками множества данных, функция ПРОЦЕНТРАНГ выполняет линейную интерполяцию и рассчитывает корректное значение процентного содержания.

Функция ПРОЦЕНТРАНГ не требует предварительной ранжировки данных, она проводит ее автоматически *в порядке убывания*.

Представим результаты, рассчитанные функцией ПРОЦЕНТРАНГ на основании исходных данных из табл. 5.1.

Формула	Результат
=ПРОЦЕНТРАНГ(C4:C8;57)	0,25
=ПРОЦЕНТРАНГ(C4:C8;48)	0,00
=ПРОЦЕНТРАНГ(C4:C8;95)	1,00
=ПРОЦЕНТРАНГ(C4:C8;60)	0,50
=ПРОЦЕНТРАНГ(C4:C8;77)	0,75
=ПРОЦЕНТРАНГ(C4:C8;55)	0,055
=ПРОЦЕНТРАНГ(C4:C8;70)	0,647

В двух последних строках рассчитаны интерполяционные значения.

Рассмотрим на примере расчета последнего значения, как функция ПРОЦЕНТРАНГ производит линейную интерполяцию (рис. 5.3).

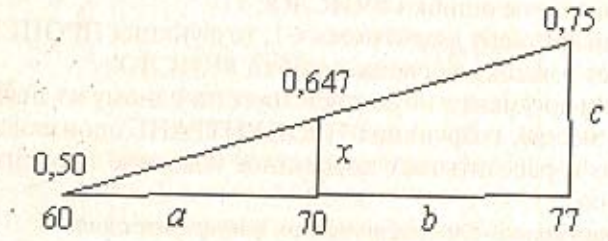

Рис. 5.3

На основании свойств подобия треугольников имеем

$$x = \frac{ac}{a+b},$$

откуда $x = \dfrac{(70-60)\cdot(0{,}75-0{,}50)}{(70-60)+(77-70)} = 0{,}147.$

Так как началом отсчета является точка 0,50, то искомое значение y определяется по формуле

$$y = 0{,}50 + x = 0{,}50 + 0{,}147 = 0{,}647.$$

ГЛАВА 6
Генерация случайных чисел

6.1.
Краткие сведения из теории статистики

Одним из фундаментальных в статистическом анализе является понятие *случайной величины*. Случайной называется переменная величина, принимающая в зависимости от случая те или иные значения с определенными вероятностями.

В практических задачах обычно используются дискретные и непрерывные случайные величины.

Дискретной случайной величиной называется такая случайная величина, множество возможных значений которой либо конечно, либо бесконечно, но счетно.

Непрерывной случайной величиной называется такая случайная величина, которая может принять любое значение из некоторого конечного или бесконечного интервала.

Чтобы дать полное математическое описание случайной величины, нужно указать множество ее значений и соответствующее случайной величине распределение вероятностей на этом множестве.

В главе 2 был рассмотрен один из основных способов задания распределения дискретной случайной величины – в виде статистического ряда распределения, который представляет собой упорядоченное распределение единиц изучаемой совокупности по определенному варьирующему признаку. При задании закона распределения непрерывной случайной величины такой способ уже

неприемлем хотя бы потому, что множество её значений бесконечно и сплошь заполняет некоторый промежуток. В этом случае не представляется возможным перечислить все значения случайной величины и их вероятности в виде таблицы (построить ряд распределения) или отметить их в системе координат (построить полигон или гистограмму распределения).

Кроме того, каждое отдельное значение непрерывной случайной величины обладает нулевой вероятностью. Однако, несмотря на данное обстоятельство, нахождение возможных значений случайной величины в различных интервалах обладает различными и отличными от нуля вероятностями. Таким образом, для непрерывной случайной величины, так же как и для дискретной, можно определить закон распределения, но несколько в ином виде, чем для дискретной. Для этого используют понятие *функции распределения случайной величины**.

Функцией распределения случайной величины X называется функция $F(x)$, задающая вероятность того, что случайная величина X принимает значение, меньшее x, т. е.

$$F(x) = P(X < x).$$

Иногда функцию $F(x)$ называют *интегральной функцией распределения*.

Функция распределения вероятностей непрерывной случайной величины дает полную вероятностную характеристику ее поведения. Однако способ задания непрерывной случайной величины с помощью функции распределения не является единственным. Ее можно задать с помощью другой функции, которая называется *дифференциальной функцией распределения* или *плотностью распределения*. В некотором смысле эта функция более удобна, чем интегральная функция $F(x)$, так как последняя не в полной мере дает представление о характере распределения случайной величины в небольшой окрестности той или иной точки числовой

* Понятие функции распределения широко используется для характеристики поведения не только непрерывных случайных величин, но и дискретных случайных величин. Для дискретной случайной величины X функция распределения имеет вид

$$F(x) = \sum_{x_i < x} P(X = x_i).$$

оси. Решить эту задачу позволяет дифференциальная функция распределения, которая является первой производной интегральной функции распределения:

$$f(x) = F'(x).$$

График дифференциальной функции распределения $f(x)$ называется кривой распределения. Кривая распределения, выражающая общую закономерность данного типа распределения, называется *теоретической кривой распределения*.

В статистике широко используются различные виды теоретических распределений – нормальное распределение, биномиальное, распределение Пуассона и др. Каждое из теоретических распределений имеет специфику и свою область применения. Чаще всего в качестве теоретического распределения используется *нормальное распределение*, занимающее особое положение в статистических исследованиях.

6.2. Справочная информация по технологии работы

Режим работы «Генерация случайных чисел» служит для формирования массива случайных чисел, распределенных по одному из заданных теоретических распределений.

В зависимости от выбранного теоретического распределения (подрежима работы) меняются и параметры диалогового окна **Генерация случайных чисел**. Общими параметрами для всех подрежимов являются:

1. *Числа переменных* – вводится число столбцов значений, которые необходимо разместить в выходном диапазоне. Если это число не введено, то все столбцы в выходном диапазоне будут заполнены.

2. *Число случайных чисел* – вводится число случайных значений, которое необходимо вывести в каждом столбце выходного диапазона. Каждое случайное значение будет помещено в строке выходного диапазона. Если число случайных чисел не будет введено, все строки выходного диапазона будут заполнены.

Примечание. Данное поле деактивизировано при модельном распределении.

3. *Распределение* – в данном раскрывающемся списке выбирается тип распределения, которое необходимо использовать для генерации случайных чисел.

4. *Случайное рассеивание* – вводится «стартовое» число для генерации определенной последовательности случайных чисел. Впоследствии это число можно снова использовать для получения той же самой последовательности случайных чисел.

Примечание. Данное поле деактивизировано при модельном распределении.

5. *Выходной интервал/Новый рабочий лист/Новая рабочая книга* – *см.* подразд. 1.1.2.

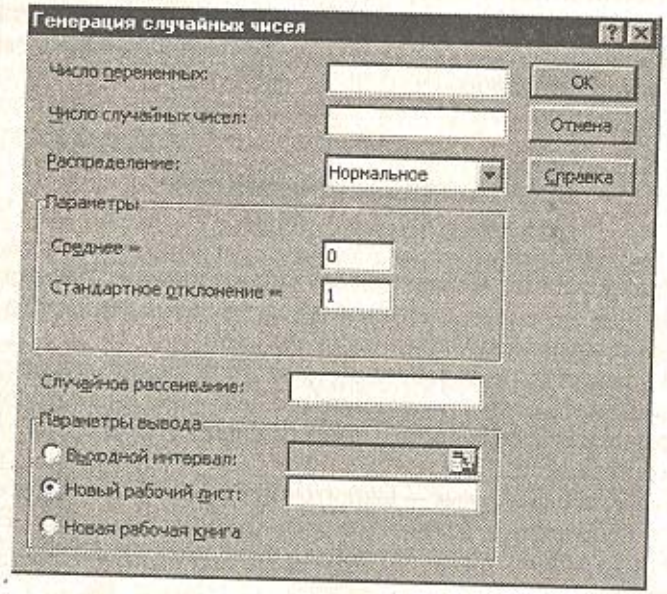

Рис. 6.1

Технология работы во всех подрежимах режима работы «Генерация случайных чисел» является одинаковой, особенность заключается только в задании параметров, характерных для конкретных распределений (как правило, они задаются в области *Параметры*).

На рис. 6.1 изображено диалоговое окно подрежима работы, предназначенного для генерации случайных чисел, распределенных по *нормальному* закону. В этом окне в области *Параметры* задаются характеристики нормального закона распределения – математическое ожидание (поле *Среднее*) и стандартное отклонение (поле *Стандартное отклонение*).

Строить графики интегральных и дифференциальных функций распределения удобно с помощью мастера диаграмм Microsoft Excel. Для этого необходимо предварительно сформировать интегральные и дифференциальные массивы значений, для чего следует воспользоваться функцией НОРМРАСП (*см.* подразд. 6.3.1), используя в качестве ее аргументов сгенерированную последовательность случайных чисел.

Графики интегральной и дифференциальной функций нормального распределения при $\bar{x} = 0$ и $\sigma^2 = 1$ показаны на рис. 6.2.

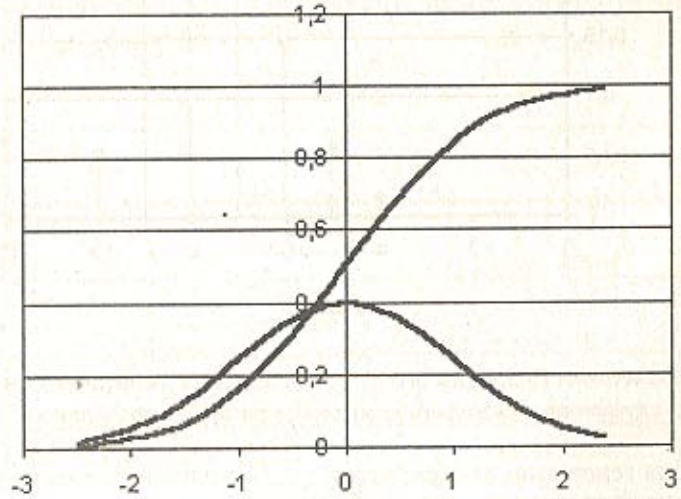

Рис. 6.2

Для генерации последовательности случайных чисел, распределенных по *биномиальному* закону, в области *Параметры* задается вероятность успеха при одном испытании (поле *Значение p*) и число испытаний (поле *Число испытаний*).

Графики биномиального распределения строятся на основе интегрального и дифференциального массивов значений, формируемых с помощью функции БИНОМРАСП (*см.* подразд. 6.4.1). Так как биномиальное распределение является дискретным, то точечные графики, построенные с помощью мастера диаграмм Microsoft Excel, необходимо дорабатывать вручную с использованием панели ***Рисование*** (нельзя использовать операцию аналитического выравнивания трендом). На рис. 6.3 изображен график дифференциальной функции биномиального распределения при $p = 0{,}75$ и $n = 10$.

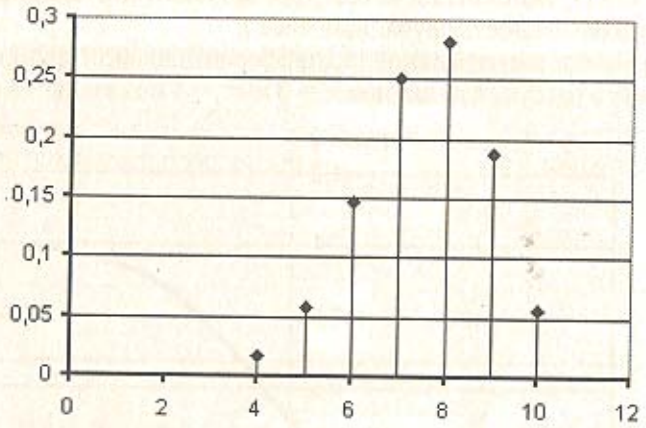

Рис. 6.3

Примечание. Подрежим работы «Распределение Бернулли» является частным случаем подрежима «Биномиальное распределение» при $n = 1$.

Для генерации последовательности случайных чисел, распределенных по *равномерному* (*прямоугольному*) закону, в области *Параметры* задаются нижняя и верхняя границы интервала, в котором будут заключены сгенерированные числа (поле *Между...*).

Понятие равномерного распределения на отрезке $[a, b]$ соответствует представлению о выборе точки из этого отрезка «наудачу». Особое значение имеет равномерное распределение на отрезке [0; 1]. Оказывается, что для имитации на ЭВМ случайных явлений самой различной природы достаточно получить на ЭВМ последовательность значений случайной величины, равномерно распределенной на отрезке [0; 1].

Интегральная и дифференциальная функции случайной величины, равномерно распределенной на отрезке [0; 1], имеют следующий вид:

$$F(x;a,b) = \begin{cases} 0 \text{ при } x \le 0; \\ x \text{ при } 0 \le x \le 1; \\ 1 \text{ при } x \ge 1; \end{cases}$$

$$f(x;a,b) = \begin{cases} 1 \text{ при } 0 \le x \le 1; \\ 0 \text{ при } x < 0 \text{ и } x > 0. \end{cases}$$

Построенный с помощью мастера диаграмм график дифференциальной функции равномерного распределения на отрезке [0; 1] изображен на рис. 6.4.

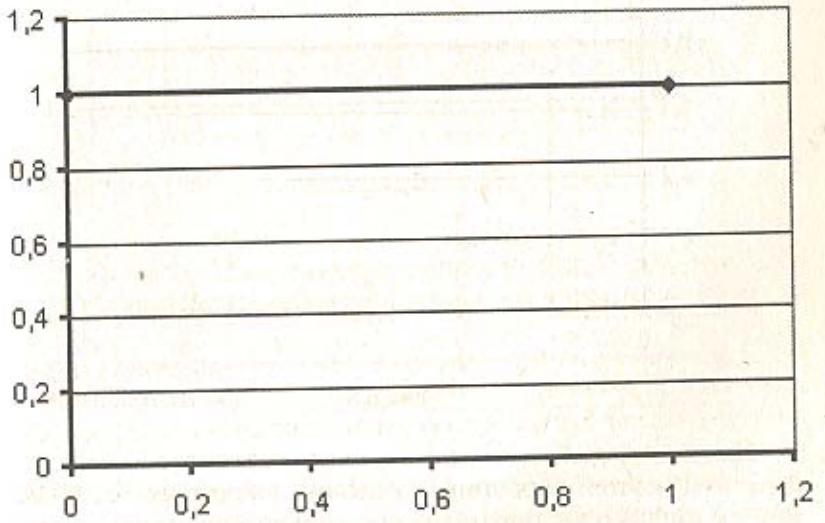

Рис. 6.4

Для генерации последовательности случайных чисел, распределенных по закону Пуассона, в области *Параметры* задается интенсивность появления событий (поле *Лямбда*).

Графики пуассоновского распределения строятся на основе интегрального и дифференциального массивов значений, формируемых с помощью функции ПУАССОН (*см.* подразд. 6.4.3). Так как распределение Пуассона является дискретным, то точечные графики, построенные с помощью мастера диаграмм Microsoft Excel, необходимо дорабатывать вручную с использованием панели ***Рисование*** (нельзя использовать операцию аналитического выравнивания трендом). На рис. 6.5 показан график дифференциальной функции распределения Пуассона при $\lambda = 0{,}8$.

Подрежим работы «Дискретное распределение» служит для генерации последовательности случайных чисел, распределенных по закону, задаваемому пользователем. В окне данного подрежима в области *Параметры* задаются значения случайной величины и соответствующие этим значениям вероятности (поле *Входной интервал значений и вероятностей*).

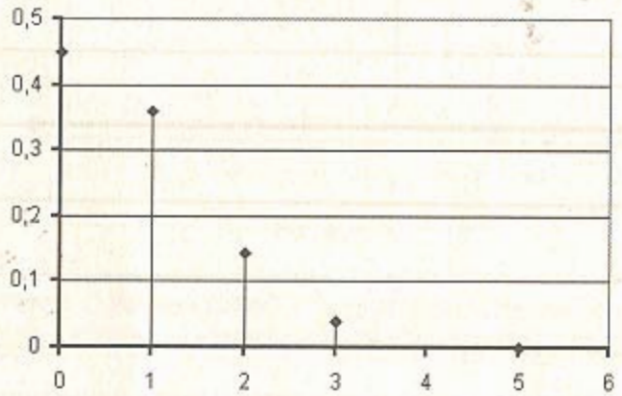

Рис. 6.5

Например, требуется смоделировать 100 подбрасываний двух игральных костей. Для этого, во-первых, на рабочем листе сформируем входную таблицу значений и вероятностей (табл. 6.1); во-вторых, зададим соответствующие параметры в диалоговом окне подрежима (рис. 6.6).

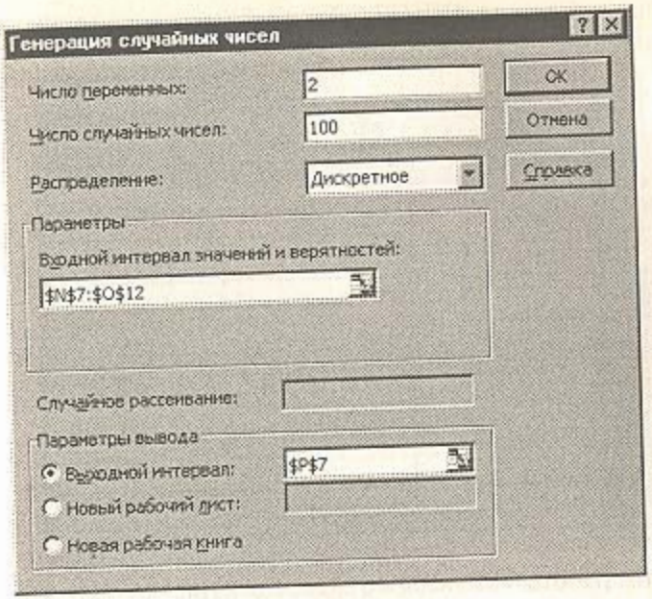

Рис. 6.6

Таблица 6.1

	N	O
6	*X*	*P* (*X*)
7	1	0,167
8	2	0,167
9	3	0,167
10	4	0,167
11	5	0,167
12	6	0,167

В результате проведенного моделирования получаем 200 значений случайной величины (100 значений для первой игральной кости в диапазоне P7:P107 и 100 значений для второй игральной кости в диапазоне Q7:Q107).

С помощью функции СЧЕТЕСЛИ посчитаем число выпавших значений для каждой игральной кости (табл. 6.2):

Таблица 6.2

	O	P	Q
108	*X*	Число выпадений	
109		Кость 1	Кость 2
110	1	18	13
111	2	15	15
112	3	13	17
113	4	26	20
114	5	14	25
115	6	14	10

Подрежим работы «Модельное распределение» служит для генерации *детерминированной последовательности* чисел в заданном интервале $[a, a_n]$ (рис. 6.7). Числа такой последовательности образуют арифметическую прогрессию, каждый член которой определяется по формуле

$$h_i = h_1 + d\,(i - 1),$$

где h_1 – первый член прогрессии (задается в поле *От...*);
d – разность прогрессии (задается в поле *Шаг*);
i – номер взятого члена.

В поле *До...* задается число, которое не может превышать последний член генерируемой прогрессии.

В подрежиме «Модельное распределение» помимо генерации чисел, образующих арифметическую прогрессию, существует возможность создания:

- нескольких одинаковых последовательностей, являющихся арифметическими прогрессиями, которые располагаются в смежных столбцах (поле *Повторяя последовательность*);
- последовательности, в которой каждое число, являющееся членом арифметической прогрессии, повторяется несколько раз (поле *Повторяя каждое число*).

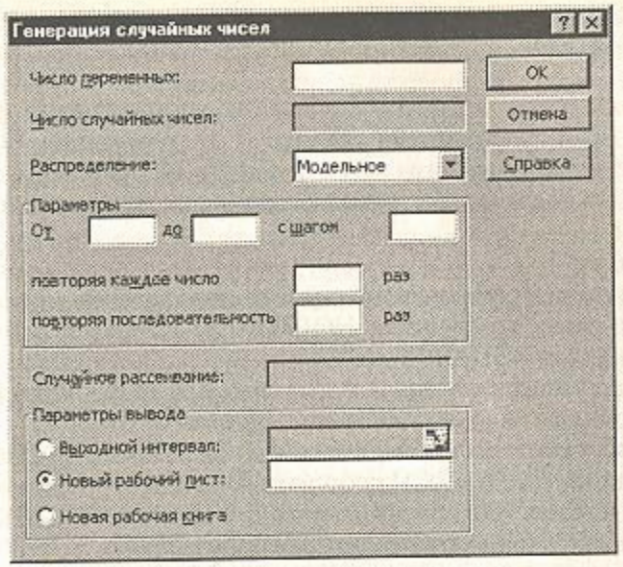

Рис. 6.7

6.3. Статистические функции непрерывных распределений

6.3.1. Функции нормального распределения

Функция НОРМРАСП

См. также НОРМОБР, НОРМСТРАСП, НОРМСТОБР, НОРМАЛИЗАЦИЯ.

Синтаксис:

НОРМРАСП (x; среднее; стандартное __ откл; интегральная)

Результат:

Рассчитывает нормальное распределение.

Аргументы:

- *x*: значение, для которого вычисляется нормальное распределение;

• *среднее*: средняя арифметическая распределения;

• *стандартное ___ откл*: стандартное отклонение распределения;

• *интегральная*: логическое значение, определяющее форму функции. Если аргумент *интегральная* = 1, то функция НОРМРАСП рассчитывает интегральную функцию распределения; если аргумент *интегральная* = 0 – дифференциальную функцию распределения.

Замечания:

• если аргумент *среднее* или аргумент *стандартное ___откл* не является числом, то функция НОРМРАСП помещает в ячейку значение ошибки #ЗНАЧ!;

• если аргумент *стандартное ___ откл* ≤ 0, то функция НОРМРАСП помещает в ячейку значение ошибки #ЧИСЛО!;

• если аргумент *среднее* = 0 и аргумент *стандартное ___откл* = =1, то функция НОРМРАСП рассчитывает стандартное нормальное распределение (*см.* описание функции НОРМСТРАСП).

Математико-статистическая интерпретация:

Нормальный закон распределения (часто называемый законом Гаусса*) имеет в статистике широкий круг приложений и занимает среди других законов распределения особое положение. Главная особенность, выделяющая нормальный закон среди других, состоит в том, что он является предельным законом, к которому приближаются другие законы распределения при весьма часто встречающихся условиях.

Доказано, что сумма достаточно большого числа независимых (или слабо зависимых) случайных величин, подчиненных каким-либо законам распределения, приближенно подчиняется нормальному закону, и это выполняется тем точнее, чем большее количество случайных величин суммируется. Основное ограничение, налагаемое на суммируемые величины, состоит в том, что

* (Gauss Carl Friedrich) Гаусс Карл Фридрих (1777–1855) – немецкий математик, внесший фундаментальный вклад также в астрономию и геодезию, иностранный чл.-корр. (1802) и иностранный почетный член (1824) Петербургской АН. Отличительными чертами творчества Гаусса являются глубокая органическая связь в его исследованиях между теоретической и прикладной математикой, необычайная широта проблематики. Работы Гаусса оказали большое влияние на развитие алгебры, теории чисел, дифференциальной геометрии, теории тяготения, классической теории электричества и магнетизма, геодезии, целых отраслей теоретической астрономии.

они все должны играть в общей сумме относительно малую роль. Если ни одна из случайно действующих величин по своему действию не окажется преобладающей над другими, то закон распределения очень близко подходит к нормальному.

Такая закономерность проявляется во многих практических случаях. Например, еще Кетле* обнаружил, что вариация в однородной группе характеризуется нормальной кривой. Если построить эмпирическую кривую распределения людей одной нации, пола и возраста по росту, весу, то она напоминает кривую Гаусса – Лапласа. Поэтому нормальное распределение часто применяется в тех случаях, когда истинный закон распределения известен, но вычисления по этому закону затруднительны, а аппроксимация его нормальным распределением допустима.

Примечание. Несмотря на широкое распространение, нормальное распределение не универсально. Если нет уверенности в его применимости, следует проверить возможность использования нормального распределения для описания случайной величины с помощью критериев согласия.

Уравнение для плотности нормального распределения имеет вид

$$f(x;\bar{x},\sigma)=\frac{1}{\sigma\sqrt{2\pi}}e^{-\frac{(x-\bar{x})^2}{2\sigma^2}},$$

а уравнение нормальной функции распределения –

$$F(x;\bar{x},\sigma)=\frac{1}{\sigma\sqrt{2\pi}}\int_{-\infty}^{x}e^{-\frac{(t-\bar{x})^2}{2\sigma^2}}dt=\Phi^*\left(\frac{x-\bar{x}}{\sigma}\right).$$

Функция НОРМРАСП использует первое уравнение, если аргумент *интегральная* = 0, и второе уравнение, если аргумент *инте-*

* (Quetelet) Кетле Ламбер Адольф Жак (1796–1874) – бельгийский ученый, социолог-позитивист. Один из создателей научной статистики, иностранный чл.-корр. Петербургской АН (1847). Установил, что некоторые массовые общественные явления (рождаемость, смертность, преступность и др.) подчиняются определенным закономерностям, применил математические методы к их изучению.

гральная = 1. Так, формула =НОРМРАСП(42;40;1,5;0) рассчитает значение 0,109, а формула =НОРМРАСП(42;40;1,5;1) – значение 0,909.

Кривая плотности нормального распределения имеет симметричный холмообразный вид (рис. 6.8).

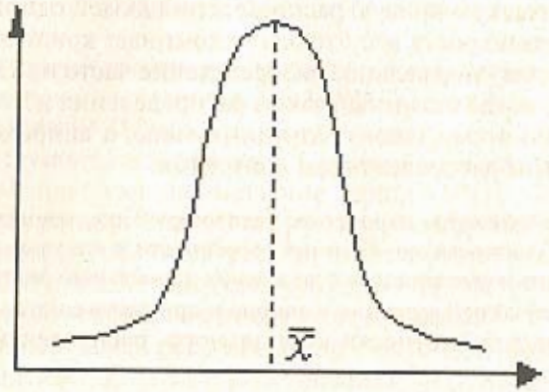

Рис. 6.8

Максимальная ордината кривой соответствует точке $x = \bar{x} = Mo = Me$. По мере удаления от этой точки плотность распределения падает, и при $x \to \pm\infty$ кривая асимптотически приближается к оси абсцисс. Изменение $\bar{x}$ при постоянстве σ приводит к смещению кривой вдоль оси абсцисс, не меняя ее формы. С увеличением σ кривая становится более пологой, с уменьшением σ – более острой. Площадь, заключенная под кривой, асимптотически приближающейся к оси абсцисс, равна единице.

Для нормального распределения выполняются следующие равенства: $\mu_1 = \mu_3 = 0$; $\mu_2 = \sigma^2$; $\mu_4 = 3\sigma^4$; $A_S = 0$; $E_k = 0$.

Весьма важной практической задачей является определение вероятности того, что случайная величина попадет на заданный интервал вещественной оси (a, b). Для нормального распределения она определяется следующей формулой:

$$P(a < x < b) = \Phi^*\left(\frac{b-\bar{x}}{\sigma}\right) - \Phi^*\left(\frac{a-\bar{x}}{\sigma}\right).$$

Пример 6.1. Для закупки и последующей продажи мужских зимних курток фирмой было проведено выборочное обследование мужского населения города в возрасте от 18 до 65 лет в целях определения его среднего роста. В результате было установлено, что средний рост $\bar{x}$ = 176 см, стандартное отклонение σ = 6 см. Необходимо определить, какой процент общего числа закупаемых курток должны составлять куртки 5-го роста (182–186 см). Предполагается, что рост мужского населения города распределен по нормальному закону.

Формула для решения задачи имеет следующий вид:

=НОРМРАСП(186;176;6;ИСТИНА)–НОРМРАСП (182;176;6;ИСТИНА) = 0,95221 – 0,84134 = 0,11086 ≈ 11%.

Таким образом, куртки 5-го роста должны составлять приблизительно 11% общего числа закупаемых курток.

Функция НОРМОБР

См. также НОРМРАСП, НОРМСТРАСП, НОРМСТОБР, НОРМАЛИЗАЦИЯ, ДОВЕРИТ.

Синтаксис:

НОРМОБР (вероятность; среднее; стандартное __ откл)

Результат:

Рассчитывает обратное нормальное распределение.

Аргументы:

- *вероятность:* вероятность, соответствующая нормальному распределению;
- *среднее*: средняя арифметическая распределения;
- *стандартное __ откл*: стандартное отклонение распределения.

Замечания:

- если какой-либо аргумент не является числом, то функция НОРМОБР помещает в ячейку значение ошибки #ЗНАЧ!;
- если аргумент *вероятность* < 0 или аргумент *вероятность* > 1, то функция НОРМОБР помещает в ячейку значение ошибки #ЧИСЛО!;
- если аргумент *стандартное__ откл* ≤ 0, то функция НОРМОБР помещает в ячейку значение ошибки #ЧИСЛО!;
- если аргумент *среднее* = 0 и аргумент *стандартное __ откл* = =1, то функция НОРМОБР использует обратное стандартное нормальное распределение (*см.* описание функции НОРМСТОБР);

• функция НОРМОБР использует для вычисления метод итераций и производит вычисления, пока не получит результат с точностью $\pm 3 \cdot 10^{-7}$. Если результат не сходится после 100 итераций, то функция помещает в ячейку значение ошибки #Н/Д.

Математико-статистическая интерпретация:

См. описание функции НОРМРАСП.

Функция обратного нормального распределения используется в ситуациях, когда известна вероятность определенного значения случайной величины и необходимо рассчитать это значение.

Например, формула =НОРМОБР(0,90879;40;1,5) рассчитывает значение 42,00001 (сравните с формулой =НОРМРАСП(42;40; 1,5;1), рассчитывающей значение 0,90879).

На практике часто встречается задача, обратная задаче вычисления вероятности попадания нормально распределенной случайной величины на участок, симметричный относительно математического ожидания $\bar{x}$. Формула для вероятности попадания случайной величины на участок, симметричный относительно математического ожидания, имеет следующий вид:

$$P(|x-\bar{x}|<l) = 2\Phi^*\left(\frac{l}{\sigma}\right) - 1,$$

где l – половина длины участка, симметричного относительно математического ожидания.

Пример 6.2. Для задачи, рассмотренной в примере 6.1, рассчитать границы интервала роста мужского населения города, вероятность попадания в который случайной величины роста составляет 0,95.

Для этого предварительно необходимо преобразовать аргументы НОРМОБР к стандартному виду, в результате чего имеем

$$l = \text{НОРМОБР}\,((P+1)/2;0;\sigma).$$

После подстановки данных получим формулу = НОРМОБР ((0,95 + 1)/2;0;6), которая рассчитает значение 11,7598. Таким образом, границы искомого интервала составят 164,24 и 187,76 см.

В качестве границ интервалов часто берутся точки, отстоящие от математического ожидания на целое число стандартных отклонений (обычно σ, 2σ, 3σ). Приведем значения вероятности попадания нормально распределенной величины в интервалы с такими границами.

Границы интервала	Вероятность
$\bar{x} - \sigma, \bar{x} + \sigma$	0,68269
$\bar{x} - 2\sigma, \bar{x} + 2\sigma$	0,95450
$\bar{x} - 3\sigma, \bar{x} + 3\sigma$	0,99730

Функция НОРМСТРАСП

См. также НОРМРАСП, НОРМОБР, НОРМСТОБР, НОРМАЛИЗАЦИЯ.

Синтаксис:

НОРМСТРАСП (z)

Результат:

Рассчитывает стандартное нормальное распределение.

Аргументы:

z: значение, для которого вычисляется стандартное нормальное распределение.

Замечания:

если аргумент *z* не является числом, то функция НОРМСТРАСП помещает в ячейку значение ошибки #ЗНАЧ!.

Математико-статистическая интерпретация:

См. описание функции НОРМРАСП.

Стандартное нормальное распределение представляет собой не что иное, как «обычное» нормальное распределение, у которого среднее равно нулю, а стандартное отклонение – единице.

Особое выделение функции стандартного нормального распределения связано с тем, что она используется при вычислении нормальных функций с другими значениями $\bar{x}$ и σ (отличными от 0 и 1 соответственно). Практически во всех учебниках по теории вероятностей и теории статистики приведены таблицы для функции стандартного нормального распределения.

Например, формула =НОРМСТРАСП((42–40)/1,5) рассчитает значение 0,90879, такое же как и формула =НОРМРАСП(42;40; 1,5;1) (*см.* описание функции НОРМРАСП).

Функция НОРМСТОБР

См. также НОРМРАСП, НОРМОБР, НОРМСТРАСП, НОРМАЛИЗАЦИЯ.

Синтаксис:

НОРМСТОБР (вероятность)

Результат:

Рассчитывает обратное стандартное нормальное распределение.

Аргументы:

вероятность: вероятность, соответствующая нормальному распределению.

Замечания:

- если аргумент *вероятность* не является числом, то функция НОРМСТОБР помещает в ячейку значение ошибки #ЗНАЧ!;
- если аргумент *вероятность* < 0 или аргумент *вероятность* > 1, то функция НОРМСТОБР помещает в ячейку значение ошибки #ЧИСЛО!;
- функция НОРМСТОБР использует для вычисления метод итераций и производит вычисления, пока не получит результат с точностью $\pm 3 \cdot 10^{-7}$. Если результат не сходится после 100 итераций, то функция помещает в ячейку значение ошибки #Н/Д.

Математико-статистическая интерпретация:

См. описание функций НОРМСТРАСП, НОРМОБР.

Функция обратного стандартного нормального распределения используется в ситуациях, когда известна вероятность определенного значения случайной величины и необходимо рассчитать это значение.

Например, формула =НОРМСТОБР(0,69146) вычисляет значение 0,5 (сравните с формулой =НОРМСТРАСП(0,5), рассчитывающей значение 0,69146). Кроме того, формула =НОРМСТОБР (0,69146) может быть заменена формулой =НОРМОБР(0,69146; 0;1), также рассчитывающей значение 0,5 (*см.* описание функции НОРМОБР).

Функция НОРМАЛИЗАЦИЯ

См. также НОРМРАСП, НОРМОБР, НОРМСТРАСП, НОРМСТОБР, ДОВЕРИТ.

Синтаксис:

НОРМАЛИЗАЦИЯ (*x*; среднее; стандартное __ откл)

Результат:

Рассчитывает нормализованное значение для нормального распределения.

Аргументы:

- *x*: нормализуемое значение;
- *среднее*: средняя арифметическая распределения;
- *стандартное __ откл.*: стандартное отклонение распределения.

Замечания:

- если аргумент *стандартное __ откл* ≤ 0, то функция НОРМАЛИЗАЦИЯ помещает в ячейку значение ошибки #ЧИСЛО!.

Математико-статистическая интерпретация:

Нормализация (нормирование) заключается в переходе от случайной величины *x* с математическим ожиданием $\bar{x}$ и дисперсией σ^2 к нормированной величине

$$t = \frac{x - \bar{x}}{\sigma},$$

получаемой в результате деления центрированной случайной величины $x - \bar{x}$ на стандартное отклонение σ. Величину *t* называют *нормированной* или *стандартизованной* случайной величиной, которая самостоятельно не применяется, а входит составной частью в выражение интегральной функции нормального распределения (*см.* описание функции НОРМРАСП).

Функцию НОРМАЛИЗАЦИЯ удобно использовать в качестве аргумента функции НОРМСТРАСП.

Например, формула =НОРМСТРАСП(НОРМАЛИЗАЦИЯ (42;40;1,5)) рассчитывает значение 0,90879, такое же как и формулы =НОРМСТРАСП((42–40)/1,5) и =НОРМРАСП(42;40;1,5;1) (*см.* описание функций НОРМСТРАСП и НОРМРАСП).

Функция ДОВЕРИТ

См. также НОРМАЛИЗАЦИЯ, НОРМОБР, НОРМРАСП, НОРМСТОБР, НОРМСТРАСП, ZТЕСТ.

Синтаксис:

ДОВЕРИТ (альфа; станд __ откл; размер)

Результат:

Рассчитывает значение предельной ошибки выборки.

Аргументы:

• *альфа*: уровень значимости, используемый для вычисления уровня надежности. Уровень надежности равняется 100 (1–*альфа*) % (например, *альфа*, равное 0,05, означает 95%-ный уровень надежности).

• *станд __ откл.*: стандартное отклонение генеральной совокупности для интервала данных, предполагается известным;

• *размер*: размер выборки.

Замечания:

• если какой-либо аргумент не является числом, то функция ДОВЕРИТ помещает в ячейку значение ошибки #ЗНАЧ!;

• если аргумент *альфа* ≤ 0 или аргумент *альфа* ≥ 1, то функция ДОВЕРИТ помещает в ячейку значение ошибки #ЧИСЛО!;

• если аргумент *станд __ откл* ≤ 0, то функция ДОВЕРИТ помещает в ячейку значение ошибки #ЧИСЛО!;

• если аргумент *размер* не целое число, то оно усекается;

• если аргумент *размер* < 1, то функция ДОВЕРИТ помещает в ячейку значение ошибки #ЧИСЛО!.

Математико-статистическая интерпретация:

Одна из основных задач выборочного исследования состоит в том, чтобы на основе характеристик выборочной совокупности получить достоверные суждения об этих характеристиках в генеральной совокупности. Возможные расхождения между характеристиками выборочной и генеральной совокупности измеряются разностью между значением характеристики в генеральной совокупности и ее значением, вычисленным по результатам выборочного наблюдения. Для средней арифметической это расхождение определяется по формуле

$$\Delta_{\tilde{x}} = |\bar{x} - \tilde{x}|.$$

Зная выборочную среднюю величину признака ($\tilde{x}$) и предельную ошибку выборки ($\Delta_{\tilde{x}}$), можно определить границы, в которых заключена генеральная средняя:

$$\tilde{x} - \Delta_{\tilde{x}} \le \bar{x} \le \tilde{x} + \Delta_{\tilde{x}}.$$

Интервал $(\tilde{x} - \Delta_{\tilde{x}}, \tilde{x} + \Delta_{\tilde{x}})$ получил название *доверительного интервала*, а величины $\tilde{x} - \Delta_{\tilde{x}}$ и $\tilde{x} + \Delta_{\tilde{x}}$ – *доверительных границ*.

Вероятность того, что случайный интервал $(\tilde{x} - \Delta_{\tilde{x}}, \tilde{x} + \Delta_{\tilde{x}})$ содержит в себе достоверную, но не известную наблюдателю характеристику $\bar{x}$, получила название *доверительной вероятности* γ. Иногда говорят, что вероятность γ характеризует надежность статистической оценки $\bar{x}$, и наряду с термином «доверительная вероятность» применяют для γ термин «надежность».

Примечание. Необходимо отметить, что в качестве аргумента функции ДОВЕРИТ используется не доверительная вероятность γ, а уровень значимости $\alpha = 1 - \gamma$, откуда $\gamma = 1 - \alpha$.

Предельная ошибка выборки $\Delta_{\tilde{x}}$ связана со средней ошибкой выборки $\mu_{\tilde{x}}$ следующим соотношением:

$$\Delta_{\tilde{x}} = t\mu_{\tilde{x}},$$

где t – коэффициент доверия (определяется в зависимости от того, с какой доверительной вероятностью нужно гарантировать результаты выборочного обследования).

Известный русский математик А. М. Ляпунов* дал выражение конкретных значений множителя t для различных значений доверительной вероятности γ в виде функции

$$\Gamma = \Phi(t) = P\{|\bar{x} - \tilde{x}| \le \Delta_{\tilde{x}}\} = \frac{1}{\sqrt{2\pi}} \int_{-t}^{t} e^{-\frac{t^2}{2}} dt.$$

На практике пользуются готовыми таблицами этой функции, которые приведены практически в каждом учебнике по теории вероятностей или теории статистики.

В Microsoft Excel для нахождения значения доверительной вероятности γ (значения функции Ф (t)) можно использовать формулу =2НОРМСТРАСП(t)– 1, а для нахождения значения коэффициента доверия t – формулу =НОРМСТОБР((γ+1)/2) (*см.* описание функции НОРМАЛИЗАЦИЯ, НОРМРАСП, НОРМОБР, НОРМСТРАСП, НОРМСТОБР).

---

* Ляпунов А. М. (1857–1918) – русский математик и механик, академик Петербургской АН (1901). Создал современную теорию устойчивости равновесия и движения механических систем с конечным числом параметров. Труды по дифференциальным уравнениям, гидродинамике, теории вероятностей.

Применение функции ДОВЕРИТ для решения практических задач рассмотрим на следующем примере.

Пример 6.3. В результате выборочного обследования жилищных условий жителей города, осуществленного на основе собственно-случайной повторной выборки, получен следующий ряд распределения (табл. 6.3). Требуется с уровнем надежности 95% определить границы интервала, в который попадет средний размер общей площади [12].

Таблица 6.3

Общая площадь, приходящаяся на 1 человека, м^2	До 5	5–10	10–15	15–20	20–25	25–30	30 и более
Число жителей	8	95	204	270	210	130	83

Рассмотрим решение задачи в среде Microsoft Excel (табл. 6.4).

Таблица 6.4

	B	C	D	E
2	Общая площадь, приходящаяся на 1 человека, м^2	Середина интервала, x	Число жителей, f	$(x-\bar{x})^2$
3	До 5,0	2,5	8	272,42
4	5,0–10,0	7,5	95	132,37
5	10,0–15,0	12,5	204	42,32
6	15,0–20,0	17,5	270	2,27
7	20,0–25,0	22,5	210	12,22
8	25,0–30,0	27,5	130	72,17
9	30,0 и более	32,5	83	182,12
10	Число жителей в выборочной совокупности, n		1000	
11	Выборочная средняя, $\bar{x}$		19,01	
12	Генеральная дисперсия, σ^2_{GEN}		51,11	

Продолжение

	B	C	D	E
2	Общая площадь, приходящаяся на 1 человека, м^2	Середина интервала, x	Число жителей, f	$(x-\bar{x})^2$
13	Генеральное стандартное отклонение, σ_{GEN}		7,15	
14	Средняя ошибка выборки, $\mu_{\bar{x}}$		0,23	
15	Коэффициент доверия, t		1,960	
16	Предельная ошибка выборки, $\Delta_{\bar{x}}$		**0,44**	
17	Нижняя граница, $\bar{x}-\Delta_{\bar{x}}$		**18,56**	
18	Верхняя граница, $\bar{x}+\Delta_{\bar{x}}$		**19,45**	
19	Предельная ошибка выборки (через ДОВЕРИТ), $\Delta_{\bar{x}}$		**0,44**	

В табл. 6.4 приведены два варианта решения задачи. Первый вариант основан на последовательном применении рассмотренных выше формул для нахождения предельной ошибки выборки $\Delta_{\bar{x}}$. Во втором варианте (более быстром) используется функция ДОВЕРИТ.

Содержимое ячеек в табл. 6.4:

- массивы B3:B9 и D3:D9 содержат исходные данные задачи;
- массив C3:C9 содержит середины рассматриваемых интервалов;
- ячейка D10 содержит формулу =СУММ (D3:D9) – рассчитывается размер выборочной совокупности n;
- ячейка D11 содержит формулу =СУММПРОИЗВ(C3:C9;D3:D9)/D10 – определяется значение выборочной средней $\bar{x}$;
- ячейка D12 содержит формулу =(СУММПРОИЗВ(E3:E9;D3:D9)/D10)*D10/(D10–1) – вычисляется значение генеральной дисперсии σ^2_{GEN} (*см.* описание функции ДИСП в подразд. 4.3);
- ячейка D13 содержит формулу =КОРЕНЬ(D12) – рассчитывается значение стандартного отклонения σ_{GEN} для генеральной совокупности;

• ячейка D14 содержит формулу =D13/КОРЕНЬ(D10) – определяется значение средней ошибки выборки $\mu_{\bar{x}}$;

• ячейка D15 содержит формулу =НОРМСТОБР((0,95+1)/2) – вычисляется значение коэффициента доверия *t* для уровня надежности 95 %;

• ячейка D16 содержит формулу =D15*D14 – рассчитывается значение предельной ошибки выборки $\Delta_{\bar{x}}$;

• ячейка D17 содержит формулу =D11–D16 – определяется нижняя граница генеральной средней $\bar{x}+\Delta_{\bar{x}}$;

• ячейка D18 содержит формулу =D11+D16 – рассчитывается верхняя граница генеральной средней $\bar{x}+\Delta_{\bar{x}}$;

• ячейка D19 содержит формулу =ДОВЕРИТ(0,05;D13; D10), демонстрирующую альтернативный вариант нахождения предельной ошибки выборки.

Таким образом, на основании проведенного выборочного обследования с уровнем надежности 95 % можно предположить, что средний размер общей площади, приходящейся на 1 человека, в целом по городу лежит в пределах от 18,56 до 19,45 м^2.

6.3.2. Функции гамма-распределения

Функция ГАММАРАСП

См. также ГАММАОБР.

Синтаксис:

ГАММАРАСП (*x*; эта; бета; интегральная)

Результат:

Рассчитывает гамма-распределение.

Аргументы:

• *x*: значение, для которого вычисляется гамма-распределение;

• *эта*: параметр распределения;

• *бета*: параметр распределения. Если *бета*=1, то функция ГАММАРАСП рассчитывает стандартное гамма-распределение;

• *интегральная*: логическое значение, определяющее форму функции. Если *интегральная*=1, то функция ГАММАРАСП рассчитывает интегральную функцию распределения; если *интегральная* = 0 – функцию плотности распределения.

Замечания:

• если аргументы *x*, *эта* или *бета* не являются числом, то функция ГАММАРАСП помещает в ячейку значение ошибки #ЗНАЧ!;

• если аргумент $x < 0$, то функция ГАММАРАСП помещает в ячейку значение ошибки #ЧИСЛО!;

• если аргумент *эта* ≤ 0 или аргумент *бета* ≤ 0, то функция ГАММАРАСП помещает в ячейку значение ошибки #ЧИСЛО!;

• если аргумент *эта*=1, то функция ГАММАРАСП рассчитывает экспоненциальное распределение;

• для целого положительного *n*, если аргументы *эта*=*n*/2, *бета*=2 и *интегральная*=1, функция ГАММАРАСП рассчитывает значение функции 1– ХИ2РАСП (*x*) с *n* степенями свободы;

• если аргумент *эта* – целое положительное число, то гамма-распределение также называется распределением Эрланга.

Математико-статистическая интерпретация:

Гамма-распределение – одна из наиболее общих статистических моделей. Она используется для описания случайных величин, ограниченных с одной стороны.

Плотность гамма-распределения имеет вид

$$f(x;\eta,\lambda)=\begin{cases}\dfrac{\lambda^{\eta}}{\Gamma(\eta)}x^{\eta-1}e^{-\lambda x}, & x\geq 0,\ \lambda>0,\ \eta>0;\\ 0 - \text{в остальных случаях,}\end{cases}$$

где $\Gamma(\eta)$ – гамма-функция* (интеграл Эйлера** 2-го рода):

$$\Gamma(\eta)=\int_0^{\infty}x^{\eta-1}e^{-x}dx.$$

Если η – целое положительное число, то $\Gamma(\eta)=(\eta-1)!$.

---

*Названия «гамма-функция» и «эйлеров интеграл», а также обозначение $\Gamma(\eta)$ предложил А. Лежандр (1814).

**(Euler Leonhard) Эйлер Леонард (1707–1783) – крупнейший математик, механик и физик. Родился в Швейцарии, в 1727 г. приехал в Россию, где работал сначала в качестве адъюнкта Петербургской АН, а затем (с 1783 г.) в качестве ее академика. Написал свыше 800 работ. Во всех физико-математических науках сделал важнейшие открытия. Много содействовал развитию русской науки.

При изменении параметра η изменяется форма кривой распределения. В частности, при $\eta \le 1$ график плотности распределения имеет вид кривой убывающей функции, а при $\eta > 1$ представляет собой одновершинную кривую с максимумом в точке $x=(\eta-1)/\lambda$. При изменении параметра λ форма распределения не изменяется, а меняется только его масштаб. Таким образом, η – параметр формы, а λ – параметр масштаба.

Разнообразие форм кривых гамма-распределения объясняет его широкое применение в качестве статистической модели. Так, опытным путем было обнаружено, что многие случайные величины, для которых невозможно теоретически обосновать применимость гамма-распределения, хорошо аппроксимируются этой статистической моделью. В числе примеров можно назвать распределение размера доходов семей и времени безотказной работы конденсатора. Кроме того, гамма-распределение часто используется в байесовском анализе как априорная модель, описывающая интенсивность некоторого процесса, когда вначале точная форма распределения неизвестна.

Наиболее широкое применение гамма-распределение нашло при описании времени, необходимого для появления ровно η независимых событий, если эти события происходят с постоянной интенсивностью λ. Этим объясняется тот факт, что гамма-распределение имеет исключительно важную роль в теории массового обслуживания, где рассматриваются задачи, связанные с ожиданием в очереди и обслуживанием клиентов. Так, например, если поставка некоторой продукции производится партиями объемом η, а заявки на продукцию поступают независимо друг от друга с постоянной интенсивностью λ единиц в неделю, то промежуток времени, за который будет реализована вся партия продукции, является случайной величиной, имеющей гамма-распределение.

Интегральная функция гамма-распределения называется *неполной гамма-функцией* и имеет следующий вид:

$$F(x;\eta,\lambda)=\begin{cases}\dfrac{\lambda^{\eta}}{\Gamma(\eta)}\displaystyle\int_0^x t^{\eta-1}e^{-\lambda t}dt, x\ge 0;\\ 0, x<0.\end{cases}$$

Неполная гамма-функция определяет вероятность того, что случайная величина, имеющая гамма-распределение с параметрами η и λ, принимает значение, меньшее x.

Применение функции ГАММАРАСП для вычисления значений гамма-распределения рассмотрим на примере 6.4. Перед этим заметим, что аргумент функции *бета* (β) является обратным по отношению к рассмотренному аргументу λ, т. е. $\lambda = 1/\beta$.

Пример 6.4. Паромная переправа осуществляет доставку контейнеров на другой берег реки. Паром отправляется в рейс, как только на него погрузят 10 контейнеров. В течение определенного времени контейнеры доставляются на паром независимо друг от друга со средней интенсивностью 6 контейнеров в час. Требуется определить вероятность того, что время между последовательными рейсами парома будет: а) менее 1 часа; б) менее 1,5 часа; в) менее 2 часов.

Для решения задачи воспользуемся функцией ГАММАРАСП, которая рассчитает следующие значения:

а) 0,084 (формула =ГАММАРАСП(1;10;1/6;1));

б) 0,413 (формула =ГАММАРАСП(1,5;10;1/6;1));

в) 0,758 (формула =ГАММАРАСП(2;10;1/6;1)).

Математическое ожидание и дисперсия гамма-распределения имеют следующий вид:

$$a(x)=\frac{\eta}{\lambda};$$

$$\sigma^2(x)=\frac{\eta}{\lambda^2}.$$

Если $\eta = 1$, то функция ГАММАРАСП рассчитывает экспоненциальное распределение (*см.* описание функции ЭКСПРАСП в подразд. 6.3.5). Если $\eta = n/2$, где n – целое положительное число, и $\beta = 2$ (или, что то же самое, $\lambda = 1/2$), функция ГАММАРАСП рассчитывает значение функции 1–ХИ2РАСП (x) с n степенями свободы (*см.* описание функции ХИ2РАСП в подразд. 6.3.7).

Функция ГАММАОБР

См. также ГАММАРАСП.

Синтаксис:

ГАММАОБР (вероятность; альфа; бета)

Результат:

Рассчитывает обратное гамма-распределение.

Аргументы:

• *вероятность:* вероятность, связанная с гамма-распределением;

• *альфа*: параметр распределения;

• *бета*: параметр распределения. Если *бета*=1, то функция ГАММАРАСП рассчитывает стандартное гамма-распределение.

Замечания:

• если какой-либо аргумент не является числом, то функция ГАММАОБР помещает в ячейку значение ошибки #ЗНАЧ!;

• если аргумент *вероятность* < 0 или аргумент *вероятность* > 1, то функция ГАММАОБР помещает в ячейку значение ошибки #ЧИСЛО!;

• если аргумент *альфа* ≤ 0 или аргумент *бета* ≤ 0, то функция ГАММАОБР помещает в ячейку значение ошибки #ЧИСЛО!;

• функция ГАММАОБР для вычисления значения использует метод итераций и производит вычисления, пока не получит результат с точностью $\pm 3\cdot10^{-7}$. Если результат не сходится после 100 итераций, то функция помещает в ячейку значение ошибки #Н/Д.

Математико-статистическая интерпретация:

См. описание функции ГАММАРАСП.

Функция обратного гамма-распределения используется в ситуациях, когда известна вероятность определенного значения случайной величины и необходимо рассчитать это значение.

Например, формула =ГАММАОБР(0,75761;10;1/6) рассчитывает значение 2 (сравните с формулой =ГАММАРАСП(2;10;1/6; 1), рассчитывающей значение 0,75761).

Пример 6.5. Для задачи, рассмотренной в примере 6.4, требуется определить время t между последовательными рейсами парома, вероятность превышения которого составляет 95 %.

Решение заключается в нахождении значения t, удовлетворяющего уравнению

$$0{,}05 = F(t;10,6) = \frac{6^{10}}{\Gamma(10)}\int_0^t t^9 e^{-6t}\,dt.$$

Для решения данного уравнения используем функцию ГАММАОБР, при этом заметим, что, как и в функции ГАММАРАСП, аргумент *бета* (β) является обратным по отношению к аргументу λ, т. е. $\lambda = 1/\beta$. Учитывая это обстоятельство, формулу нахождения t запишем в виде =ГАММАОБР(0,05;10;1/6), которая рассчитает значение 0,90. Таким образом, с вероятностью 0,95 можно предположить, что время между отправлениями парома превысит 54 мин.

Функция ГАММАНЛОГ

См. также ГАММАРАСП, ГАММАОБР.

Синтаксис:

ГАММАНЛОГ (x)

Результат:

Рассчитывает натуральный логарифм гамма-функции.

Аргументы:

x: значение, для которого вычисляется натуральный логарифм гамма-функции.

Замечания:

• если аргумент x не является числом, то функция ГАММАНЛОГ помещает в ячейку значение ошибки #ЗНАЧ!;

• если аргумент $x \le 0$, то функция ГАММАНЛОГ помещает в ячейку значение ошибки #ЧИСЛО!;

• число e, возведенное в степень ГАММАНЛОГ(i), где i – целое число, рассчитывает такой же результат, как и $(i-1)!$.

Математико-статистическая интерпретация:

См. описание функции ГАММАРАСП.

В самостоятельном виде функция ГАММАНЛОГ имеет в основном теоретическое значение, однако в комбинации с другими функциями она может использоваться в расчетах, связанных с решением практических задач.

В основе функции ГАММАНЛОГ лежит *ряд Стирлинга**:

---

*(Stirling James) Стирлинг Джеймс (1692–1770) – шотландский математик, член Лондонского королевского общества (1729). Наиболее важный труд – «Метод разностей», где Стирлинг впервые дал асимптотическое разложение логарифма гамма-распределения (т. н. *ряд Стирлинга*). Некоторые результаты Д. Стирлинга были получены также Л. Эйлером в его более общих исследованиях. *Формула Стирлинга* легко получается из ряда Стирлинга, но у самого Д. Стирлинга в явном виде не встречается.

$$\ln\Gamma(x) = x\ln x - x - \frac{1}{2}\ln x + \frac{1}{2}\ln 2\pi + \varepsilon(x),$$

где $\varepsilon(x) \to 0$ при $x \to \infty$.

Из ряда Стирлинга получается *формула Стирлинга*, позволяющая находить приближенные значения гамма-функции при больших значениях x и имеющая следующий вид:

$$\Gamma(x+1) \approx \sqrt{2\pi x}\,x^x e^{-x}, \operatorname{Re} x \to \infty.$$

где $\operatorname{Re} x$ – действительная часть числа x.

Формулой Стирлинга называется также и асимптотическое равенство, позволяющее находить приближенные значения факториалов:

$$n! \approx \sqrt{2\pi n}\,n^n e^{-n}, n \to \infty.$$

Таким образом, если x – целое положительное число, то формула Стирлинга для гамма-функции рассчитывает такое же значение, как и формула Стирлинга для факториала при $n = x$. Учитывая соотношение $\Gamma(x+1) = x\Gamma(x)$, получаем

$$\Gamma(x) \approx (x-1)!,$$

где x – целое положительное число.

Значения, рассчитываемые формулами =EXP(ГАММАНЛОГ(x)) и =ФАКТР(x–1) (при условии, что x – целое положительное число), приведены в табл. 6.5.

Таблица 6.5

Вид формулы	Значения x		
	5	10	15
EXP(ГАММАНЛОГ(x))	24,00000	362879,99992	87178291181,08069
ФАКТР(x–1)	24	362880	87178291200

6.3.3. Функции бета-распределения

Функция БЕТАРАСП

См. также БЕТАОБР.

Синтаксис:

БЕТАРАСП (x; альфа; бета; A; B)

Результат:

Рассчитывает бета-распределение.

Аргументы:

- x: значение в интервале между A и B, для которого вычисляется бета-распределение;
- *альфа*: параметр распределения;
- *бета*: параметр распределения;
- A: необязательная нижняя граница интервала изменения x;
- B: необязательная верхняя граница интервала изменения x.

Замечания:

- если какой-либо аргумент не является числом, то функция БЕТАРАСП помещает в ячейку значение ошибки #ЗНАЧ!;
- если аргумент *альфа* ≤ 0 или аргумент *бета* ≤ 0, то функция БЕТАРАСП помещает в ячейку значение ошибки #ЧИСЛО!;
- если $x < A$, или $x > B$, или $A = B$, то функция БЕТАРАСП помещает в ячейку значение ошибки #ЧИСЛО!;
- если аргументы A и B опущены, то функция БЕТАРАСП использует стандартное интегральное бета-распределение, при котором $A = 0$ и $B = 1$.

Математико-статистическая интерпретация:

Бета-распределение является одной из наиболее общих статистических моделей и используется для описания случайных величин, значения которых ограничены конечным интервалом (сравните с гамма-распределением, описывающем случайные величины, ограниченные только с одной стороны (*см.* описание функции ГАММАРАСП в подразд. 6.3.2)).

Наибольшее распространение получило стандартное бета-распределение, определенное на интервале [0; 1]. При использовании функции БЕТАРАСП это обстоятельство учитывается, если опустить в ней аргументы A и B или присвоить им значения 0 и 1 соответственно.

Плотность бета-распределения имеет вид

$$f(x;\alpha,\beta)=\begin{cases}\dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}x^{\alpha-1}(1-x)^{\beta-1} & ,0\le x\le 1,\ \alpha>0,\ \beta>0;\\ 0-\text{в остальных случаях,}\end{cases}$$

где $\dfrac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}=B(\alpha,\beta)=\int\limits_0^1 x^{\alpha-1}(1-x)^{\beta-1}dx$ – бета-функция (интеграл Эйлера 1-го рода), выраженная через гамма-функцию $\Gamma(\eta)$ (интеграл Эйлера 2-го рода);

$$\Gamma(\alpha)=\int\limits_0^\infty x^{\alpha-1}e^{-x}dx \quad \text{– гамма-функция.}$$

Функция БЕТАРАСП рассчитывает значение *интегральной функции бета-распределения*, которая также называется *неполной бета-функцией* и имеет следующий вид:

$$F(x;\alpha,\beta)=\begin{cases}0, x<0;\\ \dfrac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\int\limits_0^x t^{\alpha-1}(1-t)^{\beta-1}dt, 0\le x\le 1;\\ 1, x>0;\end{cases}$$

При различных значениях параметров α и β бета-распределение принимает различную форму:

- при $\alpha>1$ и $\beta>1$ – одновершинное с максимумом в точке $x=(\alpha-1)/(\alpha+\beta-2)$;
- при $\alpha<1$ и $\beta<1$ – *U*-образная форма;
- при $\alpha<1$ и $\beta\ge1$ – убывающая функция;
- при $\alpha\ge1$ и $\beta<1$ – *J*-образная форма;
- при $\alpha=\beta$ – симметричная форма.

Вследствие того что бета-распределение может принимать разнообразную форму, оно используется для описания большого числа реальных случайных величин, значения которых ограничены некоторым интервалом. Примерами такой случайной величины могут служить доля дефектных изделий на производственной линии, оценка продолжительности определенного этапа работы

при календарном планировании по методу PERT*. Бета-распределение используется также при байесовском** анализе в качестве исходной информации о вероятности успеха, например о вероятности того, что космический аппарат успешно выполнит определенную задачу.

Наиболее широкое применение бета-распределение получило при решении задач следующего типа. Допустим, что получены n независимых случайных наблюдений некоторого явления z с произвольной плотностью распределения. Полученные значения отсортированы в порядке возрастания. Пусть z_r и z_{n-s+1} – соответственно значения r-го наименьшего и s-го наибольшего значения. Можно показать, что доля x значений исходной совокупности, заключенных между z_r и z_{n-s+1}, имеет бета-распределение с параметрами $\alpha=n-r-s+1$ и $\beta=r+s$, т. е.

$$f(x;n-r-s+1,r+s)=\frac{\Gamma(n+1)}{\Gamma(n-r-s+1)\Gamma(r+s)}x^{n-r-s}(1-x)^{r+s-1}.$$

$$0\le x\le 1.$$

Этот результат справедлив независимо от формы распределения случайной величины z и иллюстрируется следующим примером.

Пример 6.6. Измерительный прибор, в состав которого входят фоточувствительные элементы, настроен на регистрацию минимально допустимой и максимально допустимой длины детали. Из очень большой партии случайным образом выбраны 30 деталей. Какова вероятность того, что доля деталей в партии, имеющих допустимую длину, составит: а) не менее 90 %; б) не менее 95 %; в) не менее 99 %?

---

*PERT (Program Evaluation and Review Technique) – метод оценки и пересмотра программ. Был разработан консультативной фирмой REND по заказу военно-морского министерства США для календарного планирования научно-исследовательских и опытно-конструкторских работ программы создания ракет «Поларис».

**(Bayes Thomas) Байес Томас (1702–1761) – английский математик, член Лондонского королевского общества. Основные труды относятся к теории вероятностей. В частности, Байес поставил и решил одну из основных задач элементарной теории вероятностей – *теорему Байеса* (опубликована в 1763 г.).

Из вышеизложенного следует, что доля x значений совокупности, заключенных между наибольшим и наименьшим значениями случайной выборки объемом 30 элементов, является случайной величиной, имеющей бета-распределение с параметрами $\alpha = 30-1-1+1=29$ и $\beta=1+1=2$. Следовательно, вероятность того, что доля деталей с допустимой длиной превысит 0,90, равна

$$P(x>0{,}90) = 1 - F(0{,}90;29,2) = 1 - \frac{\Gamma(29)}{\Gamma(29)\Gamma(2)} \int_0^{0{,}95} t^{29-1}(1-t)^{2-1}dt.$$

Аналогичным образом рассчитывается вероятность и для значений доли 0,95 и 0,99.

Решим задачу, воспользовавшись функцией БЕТАРАСП, которая рассчитает следующие значения:

а) 0,816 (формула =1-БЕТАРАСП(0,90;29;2));

б) 0,446 (формула =1-БЕТАРАСП(0,95;29;2));

в) 0,036 (формула =1-БЕТАРАСП(0,99;29;2)).

Математическое ожидание и дисперсия бета-распределения имеют следующий вид:

$$\alpha(x) = \frac{\alpha}{\alpha+\beta};$$

$$\sigma^2(x) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}.$$

Частными случаями бета-распределения являются равномерное, треугольное и параболическое распределения. Равномерное распределение получается при $\alpha=1$ и $\beta=1$; треугольное распределение – при $\alpha=2$ и $\beta=1$; параболическое распределение, при $\alpha=2$ и $\beta=2$. Равномерное распределение является статистической моделью, описывающей момент появления события, которое с равной вероятностью может появиться в любой момент данного интервала. Два последних распределения применяются в качестве простых аппроксимаций более сложных симметричных и асимметричных распределений. Так, параболическое распределение можно использовать как очень простую аппроксимацию нормального распределения, а треугольное распределение позволяет весьма приближенно описывать некоторые случайные величины, имеющие гамма-распределение.

Функция БЕТАОБР

См. также БЕТАРАСП.

Синтаксис:

БЕТАОБР (вероятность; альфа; бета; A; B)

Результат:

Рассчитывает обратное бета-распределение.

Аргументы:

- *вероятность*: вероятность, связанная с бета-распределением;
- *альфа*: параметр распределения;
- *бета*: параметр распределения;
- *A*: необязательная нижняя граница интервала изменения x;
- *B*: необязательная верхняя граница интервала изменения x.

Замечания:

- если какой-либо аргумент не является числом, то функция БЕТАОБР помещает в ячейку значение ошибки #ЗНАЧ!;
- если аргумент *альфа* ≤ 0 или аргумент *бета* ≤ 0, то функция БЕТАОБР помещает в ячейку значение #ЧИСЛО!;
- если аргумент *вероятность* ≤ 0 или аргумент *вероятность* > 1, то функция БЕТАОБР помещает в ячейку значение ошибки #ЧИСЛО!;
- если аргументы *A* и *B* опущены, то функция БЕТАОБР использует стандартное интегральное бета-распределение, при котором $A = 0$ и $B = 1$;
- функция БЕТАОБР для вычисления значения использует метод итераций и производит вычисления, пока не получит результат с точностью $\pm 3\cdot 10^{-7}$. Если результат не сходится после 100 итераций, то функция помещает в ячейку значение ошибки #Н/Д.

Математико-статистическая интерпретация:

См. описание функции БЕТАРАСП.

Функция обратного бета-распределения используется в ситуациях, когда известна вероятность определенного значения случайной величины и необходимо рассчитать это значение.

Например, формула =БЕТАОБР(0,55354;29;2) рассчитывает значение 0,95 (сравните с формулой =БЕТАРАСП(0,95;29;2), рассчитывающей значение 0,55354).

Пример 6.7. Для задачи, рассмотренной в примере 6.6, требуется определить долю деталей в партии, имеющих длину в пределах допустимых значений, с вероятностью не менее 0,9.

Для этого необходимо решить следующее уравнение:

$$1 - F(t;29,2) = 0,9.$$

Для решения данного уравнения используем функцию БЕТАОБР с аргументом *вероятность*=1–0,9=0,1. Формула =БЕТАОБР(1–0,9;29;2) рассчитает значение 0,876. Таким образом, с вероятностью 0,9 доля деталей в партии, имеющих длину в пределах допустимых значений, составит не менее 87,6%.

6.3.4. Функции логарифмического нормального распределения

Функция ЛОГНОРМРАСП

См. также ЛОГНОРМОБР.

Синтаксис:

ЛОГНОРМРАСП (*x*; среднее; стандартное __ откл)

Результат:

Рассчитывает логарифмическое нормальное распределение.

Аргументы:

- *x*: значение, для которого вычисляется логарифмическое нормальное распределение;
- *среднее*: средняя распределенной по нормальному закону величины ln (*x*);
- *стандартное __ откл.*: стандартное отклонение распределенной по нормальному закону величины ln(*x*).

Замечания:

- если какой-либо аргумент не является числом, то функция ЛОГНОРМРАСП помещает в ячейку значение ошибки #ЗНАЧ!;
- если аргумент $x \le 0$ или аргумент *стандартное __ откл* ≤ 0, то функция ЛОГНОРМРАСП помещает в ячейку значение ошибки #ЧИСЛО!.

Математико-статистическая интерпретация:

Логарифмическое нормальное распределение описывает случайную величину, логарифм которой распределен по нормальному закону с параметрами $\bar{x}$ и σ. Логарифмическим нормальным распределением, как правило, хорошо аппроксимируются случайные величины, которые образуются в результате умножения большого числа независимых или слабозависимых неотрицательных случайных величин, дисперсия каждой из которых мала по сравнению с дисперсией их суммы. С помощью центральной предельной теоремы можно показать, что распределение произведения *n* независимых случайных величин приближается к логарифмическому нормальному распределению, подобно тому как сумма *n* независимых случайных величин приближается к нормальному распределению.

Логарифмическое нормальное распределение применяется в самых различных областях – от экономики до биологии для описания процессов, в которых наблюдаемое значение составляет случайную долю предыдущего значения. Примерами могут служить распределение суммы личных доходов, размеров наследства, суммы банковских вкладов; распределение размеров организма, развитие которого происходит под влиянием большого числа незначительных воздействий, эффект каждого из которых пропорционален мгновенному значению размера организма. Логарифмическое нормальное распределение с хорошим приближением описывает распределение размера частиц при дроблении породы, содержание компонентов (химических соединений и минералов) в породах.

Плотность логарифмического нормального распределения имеет следующий вид:

$$f(x;\bar{x},\sigma) = \begin{cases} \dfrac{1}{\sigma x\sqrt{2\pi}} e^{\frac{(\ln x-\bar{x})^2}{2\sigma^2}}, & x>0,\ \sigma>0; \\ 0 - \text{в остальных случаях.} \end{cases}$$

Интегральная функция логарифмического нормального распределения имеет вид

$$F(x;\bar{x},\sigma) = \frac{1}{\sigma x\sqrt{2\pi}} \int_{-\infty}^{x} e^{\frac{(\ln x-\bar{x})^2}{2\sigma^2}}\, dt.$$

Заметим, что $\bar{x}$ и σ не являются параметрами, соответственно характеризующими центр распределения и его масштаб, как это

имело место в случае нормального распределения. Распределение имеет правостороннюю асимметрию, степень асимметрии возрастает с увеличением σ. При малых σ логарифмическое нормальное распределение близко к нормальному.

Рассмотрим один из типичных примеров применения логарифмического нормального распределения для решения финансово-экономических задач.

Пример 6.8. Для определения уровня жизни населения региона была проведена выборочная оценка сумм их валютных вкладов в филиалах Сбербанка и коммерческих банков. В результате обследования было установлено, что средняя сумма вклада составляет 500 долл., стандартное отклонение – 50 долл. Требуется определить, какой процент общего числа вкладов составляют вклады в размере от 2000 до 2500 долл. Предполагается, что суммы вкладов населения региона распределены по логарифмическому нормальному закону.

Для решения задачи используем функцию ЛОГНОРМРАСП. Заметим, что аргументами данной функции являются среднее и стандартное отклонение логарифма натурального случайной величины. Из условий задачи известны только среднее и стандартное отклонение самой случайной величины. Поэтому примем допущение, что среднее и стандартное отклонение натурального логарифма случайной величины приблизительно равны натуральному логарифму среднего и стандартного отклонения этой случайной величины.

С учетом принятого допущения формула для решения задачи будет иметь следующий вид:

=ЛОГНОРМРАСП(2500;LN(500);LN(50)–ЛОГНОРМРАСП(2000; LN(500);LN(50)=0,6595–0,6385=0,0211≈2 %.

Таким образом, вклады в размере от 2000 до 2500 долл. составляют приблизительно 2 % общего числа вкладов населения.

Функция ЛОГНОРМОБР

См. также ЛОГНОРМРАСП.

Синтаксис:

ЛОГНОРМОБР (вероятность; среднее; стандартное ___откл)

Результат:

Рассчитывает обратное логарифмическое нормальное распределение.

Аргументы:

- *вероятность*: вероятность, связанная с логарифмическим нормальным распределением;
- *среднее*: средняя распределенной по нормальному закону величины ln(*x*);
- *стандартное __откл*: стандартное отклонение распределенной по нормальному закону величины ln(*x*).

Замечания:

- если какой-либо аргумент не является числом, то функция ЛОГНОРМОБР помещает в ячейку значение ошибки #ЗНАЧ!;
- если аргумент *вероятность* < 0 или аргумент *вероятность* > 1, то функция ЛОГНОРМОБР помещает в ячейку значение ошибки #ЧИСЛО!;
- если аргумент *стандартное __ откл* ≤ 0, то функция ЛОГНОРМОБР помещает в ячейку значение ошибки #ЧИСЛО!.

Математико-статистическая интерпретация:

См. описание функции ЛОГНОРМРАСП.

Функция обратного логарифмического нормального распределения используется в ситуациях, когда известна вероятность определенного значения случайной величины и необходимо рассчитать это значение.

Например, формула =ЛОГНОРМОБР(0,54357;3;2) рассчитывает значение 25 (сравните с формулой =ЛОГНОРМРАСП(25;3; 2), рассчитывающей значение 0,54357).

Пример 6.9. Для задачи, рассмотренной в примере 6.8, требуется определить верхнюю границу интервала суммы валютных вкладов населения в филиалах Сбербанка и коммерческих банков региона, вероятность попадания в который случайной величины суммы вклада составляет 0,5, если нижняя граница искомого интервала равна 100 долл.

Для решения задачи используем функцию ЛОГНОРМОБР с аргументом *вероятность*=0,5+ЛОГНОРМРАСП(100;LN(500);LN (50)). Формула =ЛОГНОРМОБР(0,5+ЛОГНОРМРАСП(100;LN (500);LN(50));LN(500);LN(50)) рассчитает значение 24616,2.

Таким образом, с вероятностью 0,5 можно предположить, что валютные вклады населения будут заключены в интервале от 100 до 24616,2 долл.

6.3.5. Функции экспоненциального распределения

Функция ЭКСПРАСП

См. также ВЕЙБУЛЛ, ПУАССОН.

Синтаксис:

ЭКСПРАСП (x; лямбда; интегральная)

Результат:

Рассчитывает экспоненциальное распределение.

Аргументы:

- *x*: значение, для которого вычисляется экспоненциальное распределение;
- *лямбда*: параметр распределения;
- *интегральная*: логическое значение, определяющее форму функции. Если аргумент *интегральная*=1, то функция ЭКСПРАСП рассчитывает интегральную функцию распределения; если аргумент *интегральная*=0 – дифференциальную функцию распределения.

Замечания:

- если аргумент *x* или аргумент *лямбда* не являются числом, то функция ЭКСПРАСП помещает в ячейку значение ошибки #ЗНАЧ!;
- если аргумент $x < 0$, то функция ЭКСПРАСП помещает в ячейку значение ошибки #ЧИСЛО!;
- если аргумент *лямбда* ≤ 0, то функция ЭКСПРАСП помещает в ячейку значение ошибки #ЧИСЛО!.

Математико-статистическая интерпретация:

Экспоненциальное распределение наиболее широко используется в качестве статистической модели для времени безотказной работы. Оно играет основную роль в теории надежности, подобно тому как нормальное распределение играет основную роль в других областях. Это распределение описывает время до момента появления одного события, если события появляются независимо друг от друга с постоянной средней интенсивностью.

Плотность экспоненциального распределения имеет следующий вид:

$$f(x,\lambda) = \begin{cases} \lambda e^{-\lambda x}, & x \geq 0, \lambda > 0; \\ 0 & \text{– в остальных случаях,} \end{cases}$$

где λ – интенсивность отказов.

Интегральная функция экспоненциального распределения имеет вид

$$F(x;\lambda) = \int_0^x \lambda e^{-\lambda t} dt = 1 - e^{-\lambda x}.$$

Например, если частицы попадают в счетчик независимо друг от друга со средней интенсивностью $\lambda = 2$ частицы в секунду, то вероятность того, что частица поступит в счетчик не позже, чем через секунду после предыдущей, будет равна

$$F(1;2) = 1 - e^{-2 \cdot 1} = 0{,}865.$$

Данное значение может быть получено с помощью функции =ЭКСПРАСП(1;2;1), рассчитывающей 0,865.

Наиболее широко экспоненциальное распределение используется как статистическая модель для определения времени безотказной работы отдельных компонентов или системы, когда интенсивность отказов считается постоянной. Следует заметить, что экспоненциальное распределение более приемлемо в качестве статистической модели для определения времени безотказной работы сложной системы, даже если распределение времени безотказной работы отдельных ее компонентов не является экспоненциальным. Вместе с тем необходимо отметить, что простота теории и связанных с ней вычислений не должна создавать впечатления, будто время безотказной работы любых компонентов имеет экспоненциальное распределение. Такое допущение может быть так же ошибочным, как и допущение об универсальности нормального распределения в задачах, не связанных с испытаниями на долговечность, и даже более ошибочным, поскольку во многих случаях экспоненциальное распределение не обладает такими устойчивыми свойствами, как нормальное распределение. Справедливость принятого допущения о виде распределения можно оценить на основе критериев согласия (*см.* описание функции ХИ2РАСП в подразд. 6.3.7).

Рассмотрим один из типичных примеров применения экспоненциального распределения для решения задач из теории надежности.

Пример 6.10. Установлено, что время безотказной работы источника бесперебойного питания системы подчиняется экспоненциальному закону распределения. Среднее время между появлением двух смежных отказов равно 1200 ч. Требуется определить вероятность безотказной работы источника питания к моменту x после его включения, если: а) x=100 ч; б) x=1000 ч; в) x=1200 ч.

Очевидно, что интенсивность отказов $\lambda=1/1200$. Вероятность безотказной работы в промежутке времени $(0, x)$ определяется выражением

$$P(x;\lambda) = 1 - F(x;\lambda) = e^{-\lambda x}.$$

Решим задачу, используя функцию ЭКСПРАСП, которая рассчитает следующие значения:

а) 0,9200 (формула =1–ЭКСПРАСП(100;1/1200;1));

б) 0,4346 (формула =1–ЭКСПРАСП(1000;1/1200;1));

в) 0,3679 (формула =1–ЭКСПРАСП(1200;1/1200;1)).

Математическое ожидание и дисперсия экспоненциального распределения имеют следующий вид:

$$a(x) = \frac{1}{\lambda};$$

$$\sigma^2(x) = \frac{1}{\lambda^2}.$$

Экспоненциальное распределение является гамма-распределением с параметром $\eta=1$ (*см.* описание функции ГАММАРАСП в подразд. 6.3.2). Например, формулы =ЭКСПРАСП(5;1/10;1) и =ГАММАРАСП(5;1;10;1) рассчитывают одно и то же значение – 0,3935.

6.3.6. Функция распределения Вейбулла

Функция ВЕЙБУЛЛ

См. также ЭКСПРАСП.

Синтаксис:

ВЕЙБУЛЛ (x; альфа; бета; интегральная)

Результат:

Рассчитывает распределение Вейбулла.

Аргументы:

- *x*: значение, для которого вычисляется распределение Вейбулла;
- *альфа*: параметр распределения;
- *бета*: параметр распределения.
- *интегральная*: логическое значение, определяющее форму функции. Если аргумент *интегральная*=1, то функция ВЕЙБУЛЛ рассчитывает интегральную функцию распределения; если аргумент *интегральная*=0 – дифференциальную функцию распределения.

Замечания:

- если аргументы *x*, *альфа* или *бета* не являются числом, то функция ВЕЙБУЛЛ помещает в ячейку значение ошибки #ЗНАЧ!;
- если аргумент $x < 0$, то функция ВЕЙБУЛЛ помещает в ячейку значение ошибки #ЧИСЛО!;
- если аргумент *альфа* ≤ 0 или аргумент *бета* ≤ 0, то функция ВЕЙБУЛЛ помещает в ячейку значение ошибки #ЧИСЛО!;
- если аргумент *альфа* = 1, то функция ВЕЙБУЛЛ рассчитывает экспоненциальное распределение.

Математико-статистическая интерпретация:

Во многих случаях неадекватность экспоненциального распределения (*см.* описание функции ЭКСПРАСП в подразд. 6.3.5) как статистической модели для времени безотказной работы обусловлена ограничительным допущением о постоянстве интенсивности отказов. Следовательно, когда вероятность отказов меняется с течением времени, необходимы более общие распределения.

Одним из таких распределений, получившим широкое практическое распространение, является распределение Вейбулла*.

Распределение Вейбулла часто принимается в качестве статистической модели для определения времени безотказной работы на основе экспериментальных данных. Удовлетворительные результаты получены для электронных ламп, вакуумных приборов, реле, шарикоподшипников. Время безотказной работы некоторых видов промышленного оборудования также имеет распределение Вейбулла.

Плотность распределения Вейбулла имеет следующий вид:

$$f(x;\alpha,\beta)=\begin{cases}\dfrac{\alpha}{\beta}\left(\dfrac{x}{\beta}\right)^{\alpha-1} e^{-\left(\frac{x}{\beta}\right)^{\alpha}}, x\geq 0, \alpha>0, \beta>0;\\ 0 - \text{в остальных случаях},\end{cases}$$

где α – параметр формы распределения;

β – параметр масштаба распределения.

В зависимости от параметра α кривая плотности распределения Вейбулла принимает самые разнообразные формы. В частности, при $\alpha > 1$ распределение Вейбулла является одновершинным и интенсивность отказов возрастает с течением времени. При $\alpha < 1$ распределение Вейбулла имеет вид кривой убывающей функции и с течением времени интенсивность отказов уменьшается. При $\alpha = 1$ интенсивность отказов постоянна и распределение Вейбулла совпадает с экспоненциальным. В данном случае параметр масштаба α распределения Вейбулла равен обратному значению параметра λ экспоненциального распределения. При $\alpha = 2$ распределение Вейбулла совпадает с распределением Рэлея**.

---

*Рассматриваемое распределение (точнее, семейство распределений) названо в честь В. Вейбулла, впервые использовавшего его для аппроксимации экспериментальных данных о прочности стали на разрыв при усталостных испытаниях и предложившего методы оценки параметров распределения.

**[Rayleigt (Strutt) John William] Рэлей Рейли (до получения титула лорда Стретт Джон Уильям) (1842–1919) – английский физик, один из основоположников теории колебаний, иностранный чл.-корр. Петербургской АН (1896), член Лондонского королевского общества (1873). Основные труды по теории линейных и нелинейных колебаний. В связи с задачей сложения многих колебаний со случайными фазами получил (1880) распределение вероятностей, названное позднее *распределением Рэлея*.

Интегральная функция распределения Вейбулла имеет вид

$$F(x;\alpha,\beta)=\int_0^x \frac{\alpha}{\beta}\left(\frac{t}{\beta}\right)^{\alpha-1} e^{-\left(\frac{t}{\beta}\right)^{\alpha}} dt = 1-e^{-\left(\frac{x}{\beta}\right)^{\alpha}}.$$

Рассмотрим один из типичных примеров применения распределения Вейбулла для решения задач из теории надежности.

Пример 6.11. Установлено, что время безотказной работы вакуумного прибора подчиняется закону распределения Вейбулла с параметрами $\alpha = 2$ и $\beta = 5$ (время испытаний выражено в годах). Требуется определить вероятность появления отказа в первые x лет, если: а) $x = 1$ год; б) $x = 1{,}5$ года; в) $x = 2$ года.

Решим задачу, используя функцию ВЕЙБУЛЛ, которая рассчитает следующие значения:

а) 0,0392 (формула =ВЕЙБУЛЛ(1;2;5;1));

б) 0,0861 (формула =ВЕЙБУЛЛ(1,5;2;5;1));

в) 0,1479 (формула =ВЕЙБУЛЛ(2;2;5;1)).

Математическое ожидание и дисперсия распределения Вейбулла имеют следующий вид:

$$a(x)=\beta\Gamma\left(\frac{1}{\alpha}+1\right);$$

$$\sigma^2(x)=\beta^2\left[\Gamma\left(\frac{2}{\alpha}+1\right)-\left[\Gamma\left(\frac{1}{\alpha}+1\right)\right]^2\right],$$

где $\Gamma\left(\frac{1}{\alpha}+1\right)$ – гамма-функция.

6.3.7. Функции χ^2-распределения (распределения Пирсона)

Функция ХИ2РАСП

См. также ХИ2ОБР, ХИ2ТЕСТ, подразд. 7.1.

Синтаксис:

ХИ2РАСП (*x*; степени __ свободы)

Результат:

Рассчитывает χ^2-распределение.

Аргументы:

- *x*: значение, для которого вычисляется χ^2-распределение;
- *степени __ свободы*: число степеней свободы.

Замечания:

- если какой-либо аргумент не является числом, то функция ХИ2РАСП помещает в ячейку значение ошибки #ЗНАЧ!;
- если аргумент *x* – отрицательное число, то функция ХИ2РАСП помещает в ячейку значение ошибки #ЧИСЛО!;
- если аргумент *степени __ свободы* не целое число, то оно усекается;
- если аргумент *степени __ свободы* < 1 или аргумент *степени __ свободы* $\geq 10^{10}$, функция ХИ2РАСП помещает в ячейку значение ошибки #ЧИСЛО!.

Математико-статистическая интерпретация:

Распределением χ^2 с *k* степенями свободы называется распределение суммы квадратов *k* независимых случайных величин, каждая из которых подчинена нормальному закону с математическим ожиданием, равным нулю, и дисперсией, равной единице. Это распределение характеризуется плотностью

$$f_k(x) = \begin{cases} \dfrac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)} x^{\frac{k}{2}-1} e^{-\frac{x}{2}} & \text{при } x > 0; \\ 0 - \text{при } x \leq 0, \end{cases}$$

где $\Gamma\left(\frac{k}{2}\right)$ – гамма-функция.

Впервые χ^2-распределение было рассмотрено Р. Хельмертом* (1876) и К. Пирсоном** (1900).

*(Helmert Friedrich Robert) Хельмерт Фридрих Роберт (1843–1917) – немецкий геодезист и математик, с 1887 г. профессор Берлинского университета. Математические труды по теории ошибок; рассмотрел χ^2-распределение.

**(Pearson Karl) Пирсон Карл (1857–1936) – английский математик, биолог, философ, член Лондонского королевского общества, с 1884 г. профессор Лондонского университета. Основные труды по математической статистике (кривые Пирсона, распределение Пирсона). Разработал теорию корреляции, тесты математической статистики и критерии согласия.

Особую известность χ^2-распределение получило из-за своей тесной связи с χ^2-критерием, получившим также название критерия согласия Пирсона. Критерий χ^2 широко применяется для проверки различных статистических гипотез (*см.* подразд. 7.1), основанных на χ^2-распределении.

Определение закона распределения случайной величины на основе статистических данных состоит в том, что исследователь, опираясь на свой опыт и имеющуюся информацию, выдвигает гипотезу о теоретическом распределении и вычисляет вероятность, характеризующую ее применимость. Если эта вероятность превосходит некоторую величину, называемую уровнем значимости, то считают, что гипотеза не противоречит опытным данным и она может быть принята. Если же вероятность мала, то гипотеза отвергается и исследователь должен либо выдвинуть другую гипотезу, либо пополнить статистический материал, либо сделать и то, и другое.

Основное преимущество χ^2-критерия – его гибкость. Этот критерий можно применять для проверки допущения о любом распределении, даже не зная параметров распределения. Основной его недостаток – нечувствительность к обнаружению адекватной модели, когда число наблюдений невелико.

Критерий согласия χ^2 вычисляется по формуле

$$\chi^2 = \sum \frac{(f_Э - f_Т)^2}{f_Т},$$

где $f_Э$ и $f_Т$ – эмпирические и теоретические частоты соответственно.

В учебниках по статистике приводятся специальные таблицы, по которым с помощью величины χ^2 определяется вероятность $P(\chi^2)$. Входами в таблицу являются значения χ^2 и число степеней свободы $k = n - 1$. Данную вероятность и рассчитывает рассматриваемая статистическая функция ХИ2РАСП.

На основе $P(\chi^2)$ выносится суждение о существенности или несущественности расхождения между эмпирическим и теоретическим распределениями. При $P > 0{,}5$ считается, что эмпирическое и теоретическое распределения близки, при $P \in [0{,}2;\ 0{,}5]$ совпадение между ними удовлетворительное, в остальных случаях – недостаточное.

Пример 6.12. Вещевой службой военного округа составляется заявка на поставку обмундирования для воинских частей на основании предположения, что рост военнослужащих подчиняется нормальному закону распределения. Для проверки данного предположения было проведено исследование одной из типовых частей гарнизона. По полученным данным требуется проверить правдоподобность выдвинутой гипотезы о распределении роста военнослужащих по нормальному закону.

Исходные данные, промежуточные результаты и решение данной задачи приведены в табл. 6.6.

Таблица 6.6

	B	C	D	E	F	G	H	I	J
2	Рост в/с, см		Число в/с, $f_э$	Середина интервала, x'	$(x'-\bar{x})^2$	$f(x';\bar{x};\sigma)$	$f_т$	$f_т$ (округленные)	$\frac{(f_э-f_т)^2}{f_т}$
3	162	166	5	164	211,06	0,00323	6,47	6	0,17
4	166	170	33	168	110,84	0,01371	27,42	27	1,33
5	170	174	70	172	42,61	0,03665	73,30	73	0,12
6	174	178	132	176	6,39	0,06177	123,54	124	0,52
7	178	182	119	180	2,17	0,06565	131,29	131	1,10
8	182	186	87	184	29,94	0,04399	87,98	88	0,01
9	186	190	42	188	89,72	0,01859	37,18	37	0,68
10	190	194	12	192	181,49	0,00495	9,91	10	0,40
11			$\Sigma f_э$	$\bar{x}'$	σ			χ^2	4,33
12			500	178,53	5,89			k	7
13								$P(\chi^2)$	**0,74**

Содержимое ячеек в табл. 6.6:

- массив B3 : D10 содержит исходные данные задачи;
- ячейка D12 содержит формулу =СУММ(D3:D10) – рассчитывается общее количество обследованных военнослужащих;
- в массиве E3:E10 определяются середины интервалов роста военнослужащих (например, ячейка E9 содержит формулу =(C3+B3)/2);

- ячейка E12 содержит формулу СУММПРОИЗВ(D3:D10;E3:E10)/D12 – вычисляется средняя арифметическая роста военнослужащих;
- ячейка F12 содержит формулу =КОРЕНЬ(СУММПРОИЗВ(D3:D10;F3:F10)/D12) – рассчитывается стандартное отклонение роста военнослужащих;
- в массиве G3:G10 вычисляются значения функции плотности нормального распределения (например, ячейка G3 содержит формулу =НОРМРАСП(E3;E12;F12;0));
- в массиве H3:H10 рассчитываются теоретические частоты нормального распределения (например, ячейка H3 содержит формулу =G3*4*D12; здесь 4 – длина одного интервала роста военнослужащих (например, 166 – 162 = 4));
- в массиве I3:I10 определяются округленные теоретические частоты, рассчитанные в массиве H3:H10;
- в массиве J3:J10 вычисляются промежуточные результаты, используемые в дальнейшем для расчета критерия χ^2 (например, ячейка J3 содержит формулу =СТЕПЕНЬ(D3–I3;2)/I3);
- ячейка J11 содержит формулу =СУММ(J3:J10) – рассчитывается значение критерия χ^2;
- ячейка J12 содержит формулу =СЧЕТ(J3:J10) – 1 – определяется число степеней свободы k;
- ячейка J13 содержит формулу =ХИ2РАСП(J11;J12) – вычисляется значение вероятности $P(\chi^2)$.

Искомая вероятность $P(\chi^2) = P(4,33) = 0,74 > 0,5$, следовательно, гипотезу о том, что рост военнослужащих распределен по нормальному закону, можно считать правдоподобной.

Другой подход к решению примера 6.12, основанный на проверке попадания χ^2-критерия в критическую область, рассмотрен в описании функции ХИ2ОБР.

Функция ХИ2ОБР

См. также ХИ2РАСП, ХИ2ТЕСТ, подразд. 7.1.

Синтаксис:

ХИ2ОБР (вероятность; степени __ свободы)

Результат:

Рассчитывает обратное χ^2-распределение.

Аргументы:

• *вероятность*: вероятность, связанная с χ^2-распределением (уровень значимости α);

• *степени __ свободы*: число степеней свободы.

Замечания:

• если какой-либо аргумент не является числом, то функция ХИ2ОБР помещает в ячейку значение ошибки #ЗНАЧ!;

• если аргумент *вероятность* < 0 или аргумент *вероятность* > 1, то функция ХИ2ОБР помещает в ячейку значение ошибки #ЧИСЛО!;

• если аргумент *степени __ свободы* не целое число, то оно усекается;

• если аргумент *степени __ свободы* < 1 или аргумент *степени __ свободы* $\geq 10^{10}$, функция ХИ2ОБР помещает в ячейку значение ошибки #ЧИСЛО!;

• функция ХИ2ОБР использует метод итераций для вычисления значения и производит вычисления, пока не получит результат с точностью $\pm 3 \cdot 10^{-7}$. Если результат не сходится после 100 итераций, то функция помещает в ячейку значение ошибки #Н/Д.

Математико-статистическая интерпретация:

См. описание функции ХИ2РАСП, подразд. 7.1.

Функция ХИ2ОБР используется в ситуациях, когда известна вероятность $P(\chi^2)$ и необходимо рассчитать значение χ^2-критерия.

Например, формула =ХИ2ОБР(0,85;10) рассчитывает значение 5,57 (сравните с формулой =ХИ2РАСП(5,57;10), вычисляющей значение 0,85).

Пример 6.13. В задаче, рассмотренной в примере 6.12, требуется проверить правдоподобность выдвинутой гипотезы о распределении роста военнослужащих по нормальному закону при уровне значимости $\alpha = 0{,}1$.

Для решения задачи используем функцию ХИ2ОБР. Формула =ХИ2ОБР(0,1;7) рассчитает значение 12,02, задающее правостороннюю критическую область (12,02; + ∞). Так как $\chi^2_p = 4{,}33$ не попадает в критическую область, то гипотезу о том, что рост военнослужащих имеет нормальный закон распределения, не отвергаем.

Функция ХИ2ТЕСТ

См. также ХИ2ОБР, ХИ2РАСП, подразд. 7.1.

Синтаксис:

ХИ2ТЕСТ (фактический __интервал; ожидаемый __интервал)

Результат:

Рассчитывает значение теста на соответствие между выдвинутой гипотезой и эмпирическими данными.

Аргументы:

• *фактический __ интервал*: интервал эмпирических данных;

• *ожидаемый __ интервал*: интервал теоретических данных.

Замечания:

если аргументы *фактический __ интервал* и *ожидаемый __ интервал* имеют различное количество точек данных, то функция ХИ2ТЕСТ помещает в ячейку значение ошибки #Н/Д.

Математико-статистическая интерпретация:

См. описание функции ХИ2РАСП, подразд. 7.1.

Функцию ХИ2ТЕСТ удобно использовать для нахождения значения вероятности $P(\chi^2)$ после расчета точек теоретического распределения. Эта функция сначала рассчитывает значение критерия χ^2 и число степеней свободы k, а затем искомую вероятность $P(\chi^2)$.

Пример 6.14. В примере 6.12 после определения точек теоретического распределения (ячейки I3:I10) для нахождения значения вероятности $P(\chi^2)$ используется десять промежуточных формул (*см.* содержимое ячеек J3:J10,J11,J12). Для упрощения вычислений после нахождения точек теоретического распределения целесообразнее использовать формулу =ХИ2ТЕСТ(D3:D10;I3:I10), которая сразу рассчитает значение искомой вероятности $P(\chi^2) = 0{,}74$.

6.3.8. Функции *t*-распределения (распределения Стьюдента)

Функция СТЬЮДРАСП

См. также ДОВЕРИТ, СТЬЮДРАСПОБР, ТТЕСТ.

Синтаксис:

СТЬЮДРАСП (*x*; степени __ свободы; хвосты)

Результат:

Рассчитывает *t*-распределение (распределение Стьюдента).

Аргументы:

• *x*: значение, для которого вычисляется *t*-распределение;

• *степени __ свободы*: число степеней свободы;

• *хвосты*: число рассчитываемых хвостов распределения. Если аргумент *хвосты* = 1, то функция СТЬЮДРАСП рассчитывает одностороннее *t*-распределение; если аргумент *хвосты* = 2 – двустороннее *t*-распределение.

Замечания:

• если какой-либо аргумент не является числом, то функция СТЬЮДРАСП помещает в ячейку значение ошибки #ЗНАЧ!;

• если аргумент *степени __ свободы* < 1, то функция СТЬЮДРАСП помещает в ячейку значение ошибки #ЧИСЛО!;

• аргументы *степени __ свободы* и *хвосты* усекаются до целых чисел;

• если аргумент *хвосты* – любое значение, отличное от 1 и 2, то функция СТЬЮДРАСП помещает в ячейку значение ошибки #ЧИСЛО!.

Математико-статистическая интерпретация:

При большом числе единиц выборочной совокупности ($n > 100$) распределение случайных ошибок выборочной средней в соответствии с теоремой Ляпунова нормально или приближается к нормальному по мере увеличения числа наблюдений. Вероятность выхода ошибки за определенные пределы оценивается на основе таблиц интеграла Лапласа (*см.* описание функции ДОВЕРИТ в подразд. 6.3.1).

Однако в практике статистических исследований часто приходится сталкиваться с так называемыми *малыми выборками*, объем которых не превышает 30 ед. и может доходить до 4–5 ед.

Разработка теории малой выборки была начата в 1908 г. английским статистиком Госсетом, печатавшимся под псевдонимом Стьюдент*. Он доказал, что оценка расхождения между средней малой выборкой и генеральной средней имеет особый закон распределения, получивший название *распределения Стьюдента*. Для определения возможных пределов ошибки пользуются так называемым *t*-критерием (критерием *Стьюдента*), вычисляемым по формуле

*(Student) Стьюдент [псевдоним Уильяма Сили Госсета (William Sealy Gosset)] (1876–1937) – английский математик и статистик. Один из основоположников теории статистических оценок и проверки гипотез. Установил статистическое правило проверки гипотез (критерий Стьюдента), распределение отношения двух независимых случайных величин (распределение Стьюдента).

$$t = \frac{\bar{x} - \tilde{x}}{\mu_{мв}},$$

где $\bar{x}$ – генеральная средняя;

$\tilde{x}$ – выборочная средняя;

$\mu_{мв}$ – мера случайных колебаний выборочной средней в малой выборке.

Величина $\mu_{мв}$ определяется следующей формулой:

$$\mu_{мв} = \frac{\sigma_{VIB}}{\sqrt{n-1}},$$

где величина σ_{VIB} вычисляется на основе данных выборочного наблюдения:

$$\sigma_{VIB} = \sqrt{\frac{\sum (x_i - \tilde{x})^2}{n}}.$$

Предельная ошибка малой выборки $\Delta_{мв}$ связана со средней ошибки малой выборки $\mu_{мв}$ и коэффициентом доверия *t* (критерием Стьюдента) следующим соотношением:

$$\Delta_{мв} = t\,\mu_{мв}.$$

В данном случае величина *t* связана не с нормальным распределением, а с распределением Стьюдента, которое при небольшом объеме выборки отличается от нормального: большие величины критерия имеют здесь большую вероятность, чем при нормальном распределении.

При увеличении *n* распределение Стьюдента стремится к нормальному и при $n \to \infty$ переходит в него.

Пример 6.15. При контрольной проверке качества поставленного в торговлю маргарина получены следующие данные о содержании консерванта Е205 в 10 пробах, %: 4,3; 4,2; 3,8; 4,3; 3,7; 3,9; 4,5; 4,4; 4,0; 3,9. Какова вероятность того, что среднее содержание консерванта Е205 во всей партии не выйдет за пределы 0,1% его среднего содержания в представленных пробах?

Рассмотрим решение задачи в среде Microsoft Excel (табл. 6.7).

Таблица 6.7

	B	C
2	Содержание Е205 в пробе № 1, %	4,3
3	Содержание Е205 в пробе № 2, %	4,2
4	Содержание Е205 в пробе № 3, %	3,8
5	Содержание Е205 в пробе № 4, %	4,3
6	Содержание Е205 в пробе № 5, %	3,7
7	Содержание Е205 в пробе № 6, %	3,9
8	Содержание Е205 в пробе № 7, %	4,5
9	Содержание Е205 в пробе № 8, %	4,4
10	Содержание Е205 в пробе № 9, %	4
11	Содержание Е205 в пробе № 10, %	3,9
12	Выборочная средняя, $\tilde{x}$	4,1
13	Нижняя граница, $\tilde{x} - \Delta_{\tilde{x}}$	4,0
14	Верхняя граница, $\tilde{x} + \Delta_{\tilde{x}}$	4,2
15	Стандартное отклонение σ_{VIB}	0,261
16	Средняя ошибка выборки, $\mu_{мв}$	0,087
17	Коэффициент доверия, t	1,15
18	Доверительная вероятность, γ	**0,72**

Содержимое ячеек в табл. 6.7:

• массив C2:C11 содержит исходные данные задачи;

• ячейка C12 содержит формулу =СРЗНАЧ(C2:C11) – рассчитывается значение выборочной средней $\tilde{x}$;

• ячейка C13 содержит формулу =C12–0,1 – определяется нижняя граница генеральной средней $\tilde{x} - \Delta_{\tilde{x}}$;

• ячейка C14 содержит формулу =C12+0,1 – определяется верхняя граница генеральной средней $\tilde{x} - \Delta_{\tilde{x}}$;

• ячейка C15 содержит формулу =СТАНДОТКЛОНП(C2:C11) – вычисляется стандартное отклонение σ_{VIB};

• ячейка C16 содержит формулу =C15/КОРЕНЬ(10–1) – рассчитывается значение средней ошибки выборки $\mu_{мв}$;

• ячейка C17 содержит формулу =0,1/C16 – рассчитывает значение коэффициента доверия t (здесь величина 0,1 – значение предельной ошибки выборки $\Delta_{мв}$, заданное в условии задачи);

• ячейка C18 содержит формулу =1–СТЬЮДРАСП(C17;9; 2) – рассчитывается значение доверительной вероятности γ.

Примечание. Аргументом функции СТЬЮДРАСП является число степеней свободы $k = n - 1$. Для рассматриваемой задачи $k = 10 - 1 = 9$.

Таким образом, на основании проведенного выборочного контроля качества продукции можно заключить, что среднее содержание консерванта Е205 во всей партии будет находиться в пределах от 4,0 до 4,2% с уровнем надежности 72%.

Использование t-критерия для проверки значимости линейного коэффициента корреляции рассмотрено в подразд. 13.4.

Функция СТЬЮДРАСПОБР

См. также СТЬЮДРАСП, ТТЕСТ.

Синтаксис:

СТЬЮДРАСПОБР (вероятность; степени __ свободы)

Результат:

Рассчитывает обратное t-распределение.

Аргументы:

• *вероятность*: вероятность, соответствующая двустороннему t-распределению (уровень значимости α);

• *степени __ свободы*: число степеней свободы.

Замечания:

• если какой-либо аргумент не является числом, то функция СТЬЮДРАСПОБР помещает в ячейку значение ошибки #ЗНАЧ!;

• если аргумент *вероятность* < 0 или аргумент *вероятность* > 1, то функция СТЬЮДРАСПОБР помещает в ячейку значение ошибки #ЧИСЛО!;

• если аргумент *степени __ свободы* не целое число, то оно усекается;

• если аргумент *степени __ свободы* < 1, то функция СТЬЮДРАСПОБР помещает в ячейку значение ошибки #ЧИСЛО!;

• функция СТЬЮДРАСПОБР использует метод итераций для вычисления значения и производит вычисления, пока не получит

результат с точностью $\pm 3 \cdot 10^{-7}$. Если результат не сходится после 100 итераций, то функция помещает в ячейку значение ошибки #Н/Д.

Математико-статистическая интерпретация:

См. описание функции СТЬЮДРАСП.

Функция обратного распределения Стьюдента используется в ситуациях, когда известен уровень надежности (или уровень значимости) и необходимо рассчитать значение *t*-критерия.

Например, формула =СТЬЮДРАСПОБР(0,05;4) рассчитывает значение 2,78 (сравните с формулой =СТЬЮДРАСП(2,78;4;2), вычисляющей значение 0,05).

Пример 6.16. В задаче, рассмотренной в примере 6.15, с уровнем надежности 95 % требуется определить границы интервала, в котором находится средний процент содержания консерванта Е205 в партии маргарина.

Исходя из числа степеней свободы $k(k=n-1=10-1=9)$ и заданного уровня надежности 95 % (уровня значимости $\alpha = 0,05$) находим значение коэффициента доверия, равное 2,26 (формула =СТЬЮДРАСПОБР(0,05;9)). По формуле $\Delta_{мв} = t\ \mu_{мв}$ (= 2,26 × × 0,087) находим значение предельной ошибки малой выборки, равное 0,20 (расчет значения $\mu_{мв}$ *см.* в описании функции СТЬЮДРАСП).

Следовательно, с уровнем надежности 95 % можно предположить, что во всей партии маргарина содержание консерванта Е205 находится в пределах 4,1±0,2%, т. е. от 3,9 до 4,3 %.

6.3.9. Функции *F*-распределения (распределения Фишера)

Функция FРАСП

См. также FРАСПОБР, ФТЕСТ.

Синтаксис:

FРАСП (*x*; степени __ свободы1; степени __ свободы2)

Результат:

Рассчитывает *F*-распределение (распределение Фишера).

Аргументы:

- *x*: значение, для которого вычисляется *F*-распределение;

- *степени __ свободы1*: первое число степеней свободы *k*;
- *степени __ свободы2*: второе число степеней свободы *l*.

Замечания:

- если какой-либо аргумент не является числом, то функция FРАСП помещает в ячейку значение ошибки #ЗНАЧ!;
- если аргумент *x* – отрицательное число, то функция FРАСП помещает в ячейку значение ошибки #ЧИСЛО!;
- если аргумент *степени __ свободы1* или аргумент *степени __ свободы2* не целое число, то оно усекается;
- если аргумент *степени __ свободы1* < 1 или аргумент *степени __ свободы1* $\geq 10^{10}$, функция FРАСП помещает в ячейку значение ошибки #ЧИСЛО!;
- если аргумент *степени __свободы2* < 1 или аргумент *степени __ свободы 2* $\geq 10^{10}$, функция FРАСП помещает в ячейку значение ошибки #ЧИСЛО!.

Математико-статистическая интерпретация:

Распределение Фишера (называемое иногда распределением дисперсионного отношения) – случайная величина, равная отношению двух независимых случайных величин: величины $\chi^2(k)/k$ с распределением χ^2 и *k* степенями свободы и величины $\chi^2(l)/l$ с распределением χ^2 и *l* степенями свободы. Вводя новую случайную величину

$$F(k,l) = \frac{\chi^2(k)}{k} \cdot \frac{l}{\chi^2(l)},$$

получим для нее распределение Фишера с *k* и *l* степенями свободы.

Распределение Фишера широко используется в статистике, в частности:

- при проверке адекватности уравнений регрессии (*см.* подразд. 14.2);
- при сравнении двух дисперсий (*см.* главу 9);
- при проверке гипотезы о совпадении всех коэффициентов двух уравнений линейной регрессии.

Функция FРАСП рассчитывает значение вероятности *F*-распределения. На практике чаще применяется функция FРАСПОБР, рассчитывающая значение *F*-критерия для заданного уровня значимости α и числа степеней свободы *k* и *l* (*см.* описание функции FРАСПОБР).

Например, формула =FРАСП(9,55;2;3) рассчитывает значение 0,05 (сравните с формулой =FРАСПОБР(0,05;2;3), вычисляющей значение 9,55).

Функция FРАСПОБР

См. также FРАСП, ФТЕСТ.

Синтаксис:

FРАСПОБР (вероятность; степени __ свободы1; степени __ свободы2)

Результат:

Рассчитывает обратное *F*-распределение.

Аргументы:

• *вероятность*: вероятность, соответствующая двустороннему *F*-распределению (уровень значимости α);

• *степени __ свободы1*: первое число степеней свободы *k*;

• *степени __ свободы2*: второе число степеней свободы *l*.

Замечания:

• если какой-либо аргумент не является числом, то функция FРАСПОБР помещает в ячейку значение ошибки #ЗНАЧ!;

• если аргумент *вероятность* < 0 или аргумент *вероятность* > 1, то функция FРАСПОБР помещает в ячейку значение ошибки #ЧИСЛО!;

• если аргумент *степени __ свободы 1* или аргумент *степени __ свободы 2* не целое число, то оно усекается;

• если аргумент *степени __ свободы1* < 1 или аргумент *степени __ свободы1* $\geq 10^{10}$, функция FРАСПОБР помещает в ячейку значение ошибки #ЧИСЛО!.

• если аргумент *степени __ свободы2* < 1 или аргумент *степени __ свободы2* $\geq 10^{10}$, функция FРАСПОБР помещает в ячейку значение ошибки #ЧИСЛО!;

• функция FРАСПОБР использует метод итераций для вычисления значения и производит вычисления, пока не получит результат с точностью $\pm 3 \cdot 10^{-7}$. Если результат не сходится после 100 итераций, то функция помещает в ячейку значение ошибки #Н/Д.

Математико-статистическая интерпретация:

См. описание функции FРАСП.

Функция обратного *F*-распределения используется в ситуациях, когда известен уровень надежности (или уровень значимости) и необходимо рассчитать значение *F*-критерия.

Например, формула =FРАСПОБР(0,05;2;3) рассчитывает значение 9,55 (сравните с формулой =FРАСП(9,55;2;3), вычисляющей значение 0,05).

Пример 6.17. Требуется проверить адекватность уравнения регрессии, построенного в примере 14.1.

Проверка адекватности уравнения регрессии по *F*-критерию заключается в проверке статистической значимости коэффициента детерминации R^2 на основе формулы

$$F_p = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m},$$

где n – число наблюдений;

m – число факторов в уравнении регрессии.

Примечание. Если в уравнении регрессии свободный член $a_0 = 0$, то числитель $n-m-1$ следует увеличить на 1, т. е. он будет равен $n - m$.

Для задачи, рассмотренной в примере 14.1, n=6, m=2 и уравнение регрессии имеет вид $\hat{y}=0{,}66x_1+0{,}21x_2$. Так как в данном уравнении отсутствует свободный член a_0, то числитель $n-m-1$ следует увеличить на 1, т. е. он будет равен $n-m$=6–2=4.

Ячейка C15 (*см.* табл. 14.7) содержит значение $R^2 = 0{,}994$, отсюда формула =C15*4/(1–C15)/2 рассчитает значение $F_р = 357{,}21$ (такое же значение содержит и ячейка F22 (*см.* табл. 14.8).

Исходя из числа степеней свободы k (k=m=2) и l (l=$n-m-1$=6–2–1=3) и заданного уровня надежности 95 % (уровня значимости $\alpha = 0{,}05$) находим табличное значение *F*-критерия $F^{кр}_{пр,\alpha}$, равное 9,55 (формула =FРАСПОБР(0,05;2;3)).

Так как $F_р > F^{кр}_{пр,\alpha}$, то с уровнем надежности 95 % гипотеза $H_0: R^2 = 0$ о незначимости коэффициента детерминации отвергается, следовательно, отвергается и гипотеза о несоответствии заложенных в уравнение регрессии связей реально существующим. Таким образом, построенное уравнение регрессии по *F*-критерию Фишера является адекватным.

6.4. Статистические функции дискретных распределений

6.4.1. Функции биномиального распределения

Функция БИНОМРАСП

См. также ВЕРОЯТНОСТЬ, ОТРБИНОМРАСП, КРИТБИНОМ, ГИПЕРГЕОМЕТ.

Синтаксис:

БИНОМРАСП (число __ успехов; число __ испытаний; вероятность __ успеха; интегральная)

Результат:

Рассчитывает биномиальное распределение.

Аргументы:

- *число __ успехов*: количество успешных испытаний;
- *число __ испытаний*: число независимых испытаний;
- *вероятность __ успеха*: вероятность успеха каждого испытания;
- *интегральная:* логическое значение, определяющее форму функции. Если аргумент *интегральная* = 1, то функция БИНОМРАСП рассчитывает интегральную функцию распределения, т. е. вероятность того, что число успешных испытаний не больше значения аргумента *число __ успехов*. Если аргумент *интегральная* = 0, то рассчитывается дифференциальная функция распределения, т. е. вероятность того, что число успешных испытаний в точности равно значению аргумента *число __ успехов*.

Замечания:

- аргументы *число __ успехов* и *число __ испытаний* усекаются до целых чисел;
- если аргументы *число __ успехов*, *число __ испытаний* или *вероятность __ успеха* не являются числами, то функция БИНОМРАСП помещает в ячейку значение ошибки #ЗНАЧ!;
- если аргумент *число __ успехов* < 0 или аргумент *число __ успехов* больше аргумента *число __ испытаний*, то функция БИНОМРАСП помещает в ячейку значение ошибки #ЧИСЛО!;
- если аргумент *вероятность __ успеха* < 0 или аргумент *вероятность __ успеха* > 1, то функция БИНОМРАСП помещает в ячейку значение ошибки #ЧИСЛО!.

Математико-статистическая интерпретация:

Во многих экономических и инженерных задачах рассматриваются независимые многократно повторяемые испытания, называемые *испытаниями Бернулли**. Каждое такое испытание приводит к одному из двух возможных исходов, называемых часто *успехом* и *неудачей*, и вероятность успеха p не меняется от одного опыта к другому. Наиболее знаком пример многократного подбрасывания монеты. Если монета является геометрически правильной, то $p = 0{,}5$. Часто бывает необходимо знать вероятность появления ровно x (или не менее x) успешных исходов при n независимых испытаниях.

Согласно закону умножения независимых событий вероятность появления определенной последовательности x успешных и $n-x$ неудачных исходов в n испытаниях равна $p^x(1-p)^{n-x}$, где p– вероятность успеха при одном испытании. Из комбинаторики известно, что при n испытаниях x успешных и $n-x$ неудачных исходов могут появиться C_n^x различными одинаково возможными способами:

$$C_n^x = \binom{n}{x} = \frac{n!}{x!(n-x)!}.$$

Следовательно, согласно закону сложения взаимно исключающих событий вероятность появления ровно x успешных исходов в n независимых испытаниях определяется распределением, получившим название *биномиального* (или *распределения Бернулли*):

$$f(x;p,n) = \binom{n}{x} p^x (1-p)^{n-1}, x = \overline{0,n},$$

где p – вероятность успеха при одном испытании.

---

*(Bernoulli), семья швейцарских ученых, давшая видных математиков. *Испытания Бернулли* названы в честь Якоба Бернулли (1654–1705), выдающегося ученого, ученика и сотрудника Лейбница в разработке исчисления бесконечно малых и его приложений. Основоположник теории вероятностей, где он сформулировал и доказал теорему, носящую его имя (*теорема Бернулли*).

Свое название это распределение получило из-за связи с биномом Ньютона* $(p+q)^n$, члены разложения которого представляют соответствующие вероятности различных возможных сочетаний исходов всех отдельных событий.

Вероятность появления не более r успешных исходов в n независимых испытаниях задается интегральной функцией биномиального распределения

$$P(x \le r) = F(r;p,n) = \sum_{x=0}^{r} \binom{n}{x} p^x (1-p)^{n-x}, \qquad (6.1)$$

а вероятность появления не менее r успешных исходов в n независимых испытаниях – следующей интегральной функцией биномиального распределения:

$$P(x \ge r) = F(r;p,n) = \sum_{x=r}^{n} \binom{n}{x} p^x (1-p)^{n-x}.$$

По формуле (6.1) производит вычисления функция БИНОМРАСП, если аргумент *интегральная* = 1. В случае если аргумент *интегральная* = 0, функция БИНОМРАСП рассчитывает значение функции $f(x; p, n)$.

Биномиальное распределение лежит в основе решения известной задачи, поставленной Пепусом перед Ньютоном. Суть задачи состоит в том, что из трех человек один пытается выбросить по крайней мере одну «шестерку» при шести бросках игральной кости; второй – по крайней мере две «шестерки» при двенадцати бросках кости; третий – по крайней мере три «шестерки» при восемнадцати бросках. Каковы их относительные шансы на успех? На первый взгляд может показаться, что вероятности успеха соответственно равны 1/6, 2/12, 3/18 и что все они эквивалентны. На самом деле это не так, вероятности успеха будут различными (табл. 6.8).

*Следует заметить, что название «бином Ньютона» является вдвойне неправильным, так как, во-первых, выражение $(p+q)^n$ в общем случае не является биномом («бином» означает «двучлен»); во-вторых, разложение $(p+q)^n$ для положительных n было известно и до Ньютона. Ньютону же принадлежит смелая и необычайно плодотворная мысль распространить это разложение на случай n отрицательного и дробного.

Таблица 6.8

	B	C	D	E
2	Номер игрока	Число бросков	Число выброшенных «6»	Вероятность успеха
3	1	6	≥1	0,6651
4	2	12	≥2	0,6187
5	3	18	≥3	0,5973

Содержимое ячеек в табл. 6.8:

- ячейка E3 содержит формулу =1-БИНОМРАСП(0;C3;1/6;1);
- ячейка E4 содержит формулу =1-БИНОМРАСП(1;C4;1/6;1);
- ячейка E5 содержит формулу =1-БИНОМРАСП(2;C5;1/6;1).

Рассмотрим один из типичных примеров применения биномиального распределения для решения производственных задач.

Пример 6.18. Промышленное предприятие производит крупными партиями электрические лампочки. Отдел технического контроля из каждой партии случайным образом выбирает 100 лампочек. Партия принимается, если выборка содержит не более 3 дефектных лампочек. Какова вероятность принятия партии, если в процессе производства в среднем 0,5% лампочек дефектны?

Применительно к статистике эту задачу можно сформулировать иначе: «Какова вероятность появления не более 3 успешных исходов в 100 независимых испытаниях Бернулли, если вероятность успешного исхода при одном испытании составляет 0,005?».

Для решения задачи используем функцию = БИНОМРАСП (3; 100; 0,005; 1), которая рассчитает значение 0,9983. Таким образом, вероятность принятия партии стремится к 1.

Математическое ожидание и дисперсия биномиального распределения имеют следующий вид:

$$a(x) = np,$$

$$\sigma^2(x) = np\,(1-p).$$

Биномиальное распределение симметрично при $p = 0{,}5$. При $p \neq 0{,}5$ распределение приближается к симметричному при увеличении n; приближение будет происходить тем быстрее, чем ближе значение p к 0,5. Кроме того, при увеличении n биномиальное распределение можно аппроксимировать нормальным распределением с теми же математическим ожиданием и дисперсией, т. е. $a=np$ и $\sigma^2=np(1-p)$. Это аппроксимирующее распределение дает приемлемые результаты, если np и $n(1-p)$ не менее 5.

Функция ОТРБИНОМРАСП

См. также БИНОМРАСП.

Синтаксис:

ОТРБИНОМРАСП (число __ неудач; число __ успехов; вероятность __ успеха)

Результат:

Рассчитывает распределение Паскаля.

Аргументы:

- *число __ неудач*: количество неудачных испытаний;
- *число __ успехов*: пороговое значение числа успешных испытаний;
- *вероятность __ успеха*: вероятность успеха.

Замечания:

- аргументы *число __ неудач* и *число __ успехов* усекаются до целых чисел;
- если какой-либо аргумент не является числом, то функция ОТРБИНОМРАСП помещает в ячейку значение ошибки #ЗНАЧ!;
- если аргумент *вероятность __ успеха* < 0 или аргумент *вероятность __ успеха* > 1, то функция ОТРБИНОМРАСП помещает в ячейку значение ошибки #ЧИСЛО!;
- если выражение *число __ неудач* + *число __ успехов* − 1 ≤ 0, то функция ОТРБИНОМРАСП помещает в ячейку значение ошибки #ЧИСЛО!.

Математико-статистическая интерпретация:

Функция ОТРБИНОМРАСП рассчитывает вероятность того, что при проведении независимых испытаний Бернулли, каждое с вероятностью успеха p, до появления *ровно* s успешных исходов

произойдет x неудачных исходов (или, что то же самое, потребуется всего $x+s$ испытаний). В этом случае вероятность появления x неудачных исходов описывается *распределением Паскаля**:

$$f(x;s,p)=\binom{x+s-1}{s-1}p^s(1-p)^x,$$

где x – число неудачных исходов;
s – число успешных исходов;
p – вероятность успешного исхода.

Обобщение распределения Паскаля на случай, когда s не является целым числом и факториалы в вышеприведенной формуле заменяются гамма-функциями, называется *отрицательным биномиальным распределением***. Поэтому следует отметить, что название функции ОТРБИНОМРАСП является не совсем корректным, так как данная функция оперирует только с целочисленными аргументами x и s, т. е. рассчитывает значения распределения Паскаля.

Применение функции ОТРБИНОМРАСП для решения практических задач рассмотрим на следующих примерах.

Пример 6.19. Вероятность попадания в объект управляемой авиационной бомбы оценивается как 0,6. Для гарантированного уничтожения объекта необходимо осуществить три попадания. Какова вероятность того, что для уничтожения объекта потребуется ровно: а) 3 бомбометания; б) 4 бомбометания; в) 5 бомбометаний; г) 10 бомбометаний?

Для решения задачи используем функцию ОТРБИНОМРАСП, которая рассчитает следующие значения:

а) 0,216 (формула =ОТРБИНОМРАСП(0;3;0,6));
б) 0,259 (формула =ОТРБИНОМРАСП(1;3;0,6));

---

*(Pascal Blaise) Паскаль Блез (1623–1662) – знаменитый французский философ, писатель, математик и физик. Сформулировал одну из основных теорем проективной геометрии. Работы по арифметике, теории чисел, алгебре, теории вероятностей, теории воздушного давления.

**В некоторых источниках не проводится различие между распределением Паскаля и отрицательным биномиальным распределением.

в) 0,207 (формула =ОТРБИНОМРАСП(2;3;0,6));
г) 0,013 (формула =ОТРБИНОМРАСП(7;3;0,6)).

Вероятность того, что объект будет уничтожен не более чем при 5 бомбометаниях, оценивается как 0,216 + 0,259 + 0,207 = 0,682.

Пример 6.20. Для работы в торговом представительстве необходимо отобрать двух кандидатов, обладающих целым рядом определенных профессиональных качеств. По опыту прошлых отборов замечено, что подходящий кандидат приходится в среднем на два неподходящих. Какова вероятность того, что придется провести собеседование не более чем с пятью неподходящими кандидатами, прежде чем будут найдены два подходящих кандидата?

Для решения задачи используем функцию ОТРБИНОМРАСП, которая рассчитает следующие значения:

а) 0,111 (формула =ОТРБИНОМРАСП(0;2;1/3));
б) 0,148 (формула =ОТРБИНОМРАСП(1;2;1/3));
в) 0,148 (формула =ОТРБИНОМРАСП(2;2;1/3));
г) 0,132 (формула =ОТРБИНОМРАСП(3;2;1/3));
д) 0,110 (формула =ОТРБИНОМРАСП(4;2;1/3));
е) 0,088 (формула =ОТРБИНОМРАСП(5;2;1/3)).

Вероятность того, что придется провести собеседование не более чем с пятью неподходящими кандидатами, прежде чем будут найдены два подходящих, составляет 0,737 (0,111 + 0,148 + 0,148 + 0,132 + 0,110 + 0,088 = 0,737).

Математическое ожидание и дисперсия распределения Паскаля определяются следующими выражениями:

$$a(x) = \frac{s(1-p)}{p};$$

$$\sigma^2(x) = \frac{s(1-p)}{p^2}.$$

Функция КРИТБИНОМ

См. также БИНОМРАСП, ОТРБИНОМРАСП.

Синтаксис:

КРИТБИНОМ (число __ испытаний; вероятность __успеха; альфа)

Результат:

Рассчитывает наименьшее значение, для которого интегральное биномиальное распределение больше или равно заданному критерию.

Аргументы:

- *число __ испытаний*: число испытаний Бернулли;
- *вероятность __ успеха*: вероятность успеха в каждом испытании;
- *альфа*: значение критерия.

Замечания:

- если какой-либо аргумент не является числом, то функция КРИТБИНОМ помещает в ячейку значение ошибки #ЗНАЧ!;
- если аргумент *число __ испытаний* не целое число, то оно усекается;
- если аргумент *число __ испытаний* < 0, то функция КРИТБИНОМ помещает в ячейку значение ошибки #ЧИСЛО!;
- если аргумент *вероятность __ успеха* < 0 или аргумент *вероятность __ успеха* > 1, то функция КРИТБИНОМ помещает в ячейку значение ошибки #ЧИСЛО!;
- если аргумент *альфа* < 0 или аргумент *альфа* > 1, то функция КРИТБИНОМ помещает в ячейку значение ошибки #ЧИСЛО!.

Математико-статистическая интерпретация:

См. описание функции БИНОМРАСП.

Функция КРИТБИНОМ является обратной по отношению к функции БИНОМРАСП и рассчитывает наименьшее значение, для которого интегральное биномиальное распределение больше или равно заданному критерию. Эта функция наиболее часто используется в приложениях, связанных с контролем качества продукции.

Пример. 6.21. По исходным данным примера 6.18 (за исключением числа дефектных лампочек в выборке) требуется определить наибольшее допустимое число дефектных лампочек в выборке, при котором вероятность принятия партии составит: а) 0,9; б) 0,95; в) 0,99.

Для решения задачи используем функцию КРИТБИНОМ, которая рассчитает следующие значения:

а) 1 (формула =КРИТБИНОМ(100;0,005;0,90));
б) 2 (формула =КРИТБИНОМ(100;0,005;0,95));
в) 3 (формула =КРИТБИНОМ(100;0,005;0,99)).

Из полученных результатов видно, что при ограничении «не более 1 дефектной лампочки в выборке» вероятность принятия

партии будет лежать в интервале от 0,9 до 0,95; при ограничении «не более 2 дефектных лампочек в выборке» – в интервале от 0,95 до 0,99; при ограничении «не более 3 дефектных лампочек в выборке» – в интервале от 0,99 до некоторого значения, которое можно рассчитать аналогичным образом. Точные значения вероятности принятия партии можно вычислить с помощью функции БИНОМРАСП:

- формула =БИНОМРАСП(1;100;0,005;ИСТИНА) рассчитает значение 0,910;
- формула =БИНОМРАСП(2;100;0,005;ИСТИНА) вычислит значение 0,986;
- формула =БИНОМРАСП(3;100;0,005;ИСТИНА) рассчитает значение 0,998.

Функция ПЕРЕСТ

См. также БИНОМРАСП.

Синтаксис:

ПЕРЕСТ (число; число __ выбранных)

Результат:

Рассчитывает количество перестановок для заданного числа объектов, которые выбираются из общего числа объектов.

Аргументы:

- *число*: целое число, задающее количество объектов;
- *число __ выбранных*: целое число, задающее количество объектов в каждой перестановке.

Замечания:

- аргументы усекаются до целых чисел;
- если какой-либо аргумент не является числом, то функция ПЕРЕСТ помещает в ячейку значение ошибки #ЗНАЧ!;
- если аргумент *число* ≤ 0 или аргумент *число __ выбранных* < 0, то функция ПЕРЕСТ помещает в ячейку значение ошибки #ЧИСЛО!;
- если аргумент *число* меньше аргумента *число __ выбранных*, то функция ПЕРЕСТ помещает в ячейку значение ошибки #ЧИСЛО!.

Математико-статистическая интерпретация:

Возьмем m различных элементов $a_1, a_2, \ldots, a_n$. Будем переставлять эти элементы всевозможными способами, оставляя неизменными их число и меняя лишь их порядок. Каждая из получившихся таким образом комбинаций (в том числе и первоначальная) носит название *перестановки*. Общее число перестановок из m элементов обозначается P_m. Это число равно произведению всех целых чисел от 1 (или, что то же самое, от 2) до m включительно:

$$P_m = 1 \cdot 2 \cdot 3 \ldots \cdot (m-1)m = m!.$$

Применение функции ПЕРЕСТ рассмотрим на следующем примере.

Пример 6.22. Сколькими способами можно распределить пять должностей между пятью лицами, отобранными в качестве кандидатов в отдел ценных бумаг банка? Если составить в некотором порядке список должностей и против каждой должности писать фамилию кандидатов, то каждому распределению отвечает некоторая перестановка. Общее число таких перестановок

$$P_5 = 1 \cdot 2 \cdot 3 \cdot 4 \cdot 5 = 120.$$

Для решения рассмотренной и подобных задач можно использовать функцию ПЕРЕСТ с равными значениями аргументов (*число = число __ выбранных*), формула =ПЕРЕСТ(5;5) рассчитает значение 120.

Более сложными являются задачи, когда число распределяемых элементов больше числа позиций, по которым они распределяются (для функции ПЕРЕСТ выполняется неравенство *число > число __ выбранных*). В этом случае общее число перестановок определяется следующим образом:

$$P_{m,n} = \frac{n!}{(n-m)!}.$$

Заметим, что при $n = m$, $P_{m,n} = P_m = P_n = n!$, так как $0! = 1$.

Допустим, что в качестве кандидатов на должности в отдел отобраны не пять, а шесть человек. Тогда число возможных перестановок (число возможных комбинаций распределения по должностям) составит 720 комбинаций (формула =ПЕРЕСТ(6;5)).

Функция ВЕРОЯТНОСТЬ

См. также БИНОМРАСП, КРИТБИНОМ.

Синтаксис:

ВЕРОЯТНОСТЬ(*x* __ интервал; интервал __ вероятностей; нижний __ предел; верхний __ предел)

Результат:

Рассчитывает вероятность того, что значения из интервала находятся внутри заданных границ.

Аргументы:

• *x __ интервал*: интервал числовых значений *x*, с которыми связаны вероятности;

• *интервал __ вероятностей*: множество вероятностей, соответствующих значениям в аргументе *x __ интервал*.

• *нижний __ предел*: нижняя граница значения, для которого требуется вычислить вероятность;

• *верхний __предел*: необязательная верхняя граница значения, для которого требуется вычислить вероятность.

Замечания:

• если какое-либо значение в аргументе *интервал __вероятностей* ≤ 0 или какое-либо значение в аргументе *интервал __вероятностей* > 1, то функция ВЕРОЯТНОСТЬ помещает в ячейку значение ошибки #ЧИСЛО!;

• если сумма значений в аргументе *интервал __ вероятностей* ≠1, то функция ВЕРОЯТНОСТЬ помещает в ячейку значение ошибки #ЧИСЛО!;

• если аргументы *x __ интервал* и *интервал __ вероятностей* содержат различное количество точек данных, то функция ВЕРОЯТНОСТЬ помещает в ячейку значение ошибки #Н/Д;

• если аргумент *верхний __ предел* не задан, то функция ВЕРОЯТНОСТЬ рассчитывает вероятность значения аргумента *нижний __ предел*.

Математико-статистическая интерпретация:

В основе применения функции ВЕРОЯТНОСТЬ лежит *теорема сложения вероятностей*, которая формулируется следующим образом.

Теорема. Вероятность суммы несовместных событий равна сумме вероятностей этих событий:

$$P\left(\sum_{i=1}^{n} A_i\right) = \sum_{i=1}^{n} P(A_i).$$

Из теоремы сложения вероятностей вытекает следующее следствие.

Следствие. Если события A_1, A_2, ..., A_n образуют полную группу несовместных событий, то сумма их вероятностей равна 1:

$$\sum_{i=1}^{n} P(A_i) = 1.$$

Теорема сложения вероятностей и ее следствие обусловливают математическую интерпретацию функции ВЕРОЯТНОСТЬ, применение которой рассмотрим на следующем примере.

Пример 6.23. В новогодней лотерее организации разыгрывается 1000 билетов. Из них падают выигрыши: на один билет – 500 руб.; на 10 билетов – по 100 руб.; на 50 билетов – по 20 руб.; на 100 билетов – по 5 руб.; остальные билеты невыигрышные. Сотрудник организации может взять только один билет. Найти вероятность выиграть: а) 20 руб.; б) не менее 20 руб.; в) не менее 100 руб.; г) от 5 до 100 руб. включительно.

Исходные данные и результат решения задачи с помощью функции ВЕРОЯТНОСТЬ приведены в табл. 6.9.

Содержимое ячеек в табл. 6.9:

• ячейки D3:D7 содержат соответственно формулы =B3/СУММ(B3:B7)...=B7/СУММ(B3:B7);

• ячейка D8 содержит формулу =ВЕРОЯТНОСТЬ(C3:C7;D3::D7;20);

• ячейка D9 содержит формулу =ВЕРОЯТНОСТЬ(C3:C7;D3::D7;20;500);

• ячейка D10 содержит формулу =ВЕРОЯТНОСТЬ(C3:C7;D3::D7;100;500);

• ячейка D11 содержит формулу =ВЕРОЯТНОСТЬ(C3:C7;D3::D7;5;100).

Таблица 6.9

	B	C	D
2	Количество лотерейных билетов	Размер выигрыша, руб.	Вероятность выигрыша
3	1	500	0,001
4	10	100	0,01
5	50	20	0,05
6	100	5	0,1
7	839	0	0,839
8	20		**0,05**
9		Не менее 20	**0,061**
10		Не менее 100	**0,011**
11		От 5 до 100 включительно	**0,16**

6.4.2. Функции гипергеометрического распределения

Функция ГИПЕРГЕОМЕТ

См. также БИНОМРАСП, ОТРБИНОМРАСП.

Синтаксис:

ГИПЕРГЕОМЕТ (число __ успехов __ в __ выборке; размер__ выборки; число __ успехов __ в __ совокупности; размер __ совокупности)

Результат:

Рассчитывает гипергеометрическое распределение.

Аргументы:

- *число __ успехов __ в __ выборке:* число успешных испытаний в выборке;
- *размер __ выборки:* размер выборки;
- *число __ успехов __ в __ совокупности:* число успешных испытаний в генеральной совокупности;

- *размер __ совокупности:* размер генеральной совокупности.

Замечания:

- все аргументы усекаются до целых чисел;
- если какой-либо аргумент не является числом, то функция ГИПЕРГЕОМЕТ помещает в ячейку значение ошибки #ЗНАЧ!;
- если аргумент *число __ успехов __ в __ выборке* < 0 или аргумент *число __ успехов __ в __ выборке* больше какого-либо из аргументов *размер __ выборки* и *число __ успехов __ в __ совокупности*, то функция ГИПЕРГЕОМЕТ помещает в ячейку значение ошибки #ЧИСЛО!;
- если аргумент *размер __ выборки* < 0 или аргумент *размер __ выборки* больше аргумента *размер __ совокупности*, то функция ГИПЕРГЕОМЕТ помещает в ячейку значение ошибки #ЧИСЛО!;
- если аргумент *число __ успехов __ в __ совокупности* < 0 или аргумент *число __ успехов __ в __ совокупности* больше аргумента *размер __ совокупности*, то функция ГИПЕРГЕОМЕТ помещает в ячейку значение ошибки #ЧИСЛО!;
- если аргумент *размер __ совокупности* < 0, то функция ГИПЕРГЕОМЕТ помещает в ячейку значение ошибки #ЧИСЛО!.

Математико-статистическая интерпретация:

Гипергеометрическое распределение описывает вероятность появления ровно x успешных исходов в n испытаниях, когда значение n не мало по сравнению с объемом совокупности N. Это распределение часто находит применение в задачах, когда выборка берется из небольших партий продукции. Вероятность того, что из n изделий, выбранных случайным образом из партии объемом N, ровно x являются дефектными, имеет *гипергеометрическое распределение*.

Выбрать n элементов из N можно C_N^n различными способами, каждый из которых одинаково возможен. Аналогично x из k дефектных изделий можно выбрать C_k^x различными способами. Кроме того, для каждой такой комбинации имеют место

$$\binom{N-k}{n-x}=\frac{(N-k)!}{(n-x)!\cdot[(N-k)!-(n-x)!]}$$

способов выбора $n-x$ изделий из числа $N-k$ исправных, следовательно, общее число способов выбора x дефектных изделий и $N-k$ исправных равно

$$\binom{k}{x}\binom{N-k}{n-x}=\frac{k!}{x!(k-x)!}\cdot\frac{(n-k)!}{(n-x)!\cdot[(N-k)!-(n-x)!]}.$$

А поскольку мы имеем дело с равновозможными событиями, то вероятность появления события «выбор x из k дефектных изделий и $n-x$ из $N-k$ исправных изделий» определяется следующим выражением, задающим гипергеометрическое распределение:

$$f(x;N,n,k)=\frac{\binom{k}{x}\binom{N-k}{n-x}}{\binom{N}{n}},\quad x\le k,\ n-x\le N-k,$$

где x – число дефектных изделий в выборе;
k – число дефектных изделий в генеральной совокупности;
n – объем выборки;
N – объем генеральной совокупности.

Аргументы функции $f(x; N, n, k)$ эквивалентны следующим аргументам функции ГИПЕРГЕОМЕТ:

- *x = число __ успехов __ в __ выборке*;
- *N = размер __ совокупности*;
- *n = размер __ выборки*;
- *k = число __ успехов __ в __ совокупности.*

Рассмотрим один из типичных примеров применения гипергеометрического распределения для решения задач производственного контроля качества продукции.

Пример 6.24. Из партии, содержащей 30 специальных высоконадежных электронных ламп, случайным образом выбираются и подвергаются испытаниям на долговечность шесть ламп. Если в процессе испытаний ни одна лампа не выйдет из строя или выйдет из строя только одна лампа, то партия принимается. В противном случае вся партия бракуется. Какова вероятность того, что партия будет принята, если из 30 ламп четыре являются дефектными?

Партия будет принята, если взятая выборка не содержит дефектных ламп или содержит одну дефектную лампу. Соответствующая вероятность определяется выражением $f(0;30;6;4) + f(1;30;6;4)$, которое рассчитаем с помощью функции ГИПЕРГЕОМЕТ:

=ГИПЕРГЕОМЕТ(0;6;4;30)+ГИПЕРГЕОМЕТ(0;6;4;30)=0,831.

Таким образом, с вероятностью 0,831 данная партия будет принята.

Математическое ожидание и дисперсия гипергеометрического распределения имеют следующий вид:

$$a(x)=\frac{nk}{N};$$

$$\sigma^2(x)=\frac{nk(N-k)(N-n)}{N^2(N-1)}.$$

При уменьшении отношения n/N гипергеометрическое распределение стремится к биномиальному распределению с параметрами n и $p = k/N$. Если в примере 6.24 использовать биномиальное распределение с параметрами $p = 4/30 = 0{,}133$, то вероятность принятия партии будет равна не 0,831, а 0,815 (рассчитывается по формуле =БИНОМРАСП(1;6;4/30;1)). Поскольку в данном примере выборка составляет 20 % совокупности, то нельзя ожидать, что биномиальное распределение обеспечит очень хорошую аппроксимацию.

6.4.3. Функции распределения Пуассона

Функция ПУАССОН

См. также БИНОМРАСП, ЭКСПРАСП.

Синтаксис:

ПУАССОН (*x*; среднее; интегральная)

Результат:

Рассчитывает распределение Пуассона.

Аргументы:

- *x*: количество событий;
- *среднее*: интенсивность появления событий;
- *интегральная*: логическое значение, определяющее форму функции. Если аргумент *интегральная* = 1, то функция ПУАССОН рассчитывает интегральную функцию распределения; если аргумент *интегральная* = 0 – дифференциальную функцию распределения.

Замечания:

- аргумент *x* усекается до целого числа;
- если аргумент *x* или аргумент *среднее* не является числом, то функция ПУАССОН помещает в ячейку значение ошибки #ЗНАЧ!;
- если аргумент $x \le 0$, функция ПУАССОН помещает в ячейку значение ошибки #ЧИСЛО!;
- если аргумент *среднее* ≤ 0, то функция ПУАССОН помещает в ячейку значение ошибки #ЧИСЛО!.

Математико-статистическая интерпретация:

Одним из наиболее распространенных дискретных распределений является *распределение Пуассона**, которое описывает число событий, происходящих в одинаковых промежутках времени или на одинаковых отрезках пространства при условии, что события происходят независимо друг от друга с постоянной средней интенсивностью λ.

Плотность распределения Пуассона (вероятность появления ровно *x* событий в определенном промежутке времени) имеет следующий вид:

$$f(x;\lambda) = \frac{\lambda^x}{x!} e^{-\lambda}.$$

*(Poisson Simeon Denis) Пуассон Симеон Дени (1781–1840) – французский механик, физик, математик, иностранный почетный член Петербургской АН (1826), член Парижской АН (1812). Основные труды по теоретической и небесной механике, математике и математической физике. В теории вероятностей Пуассон доказал частный случай закона больших чисел и одну из предельных теорем (*теорема Пуассона, распределение Пуассона*).

Во временной области пуассоновское распределение используется как статистическая модель для числа альфа-частиц, испускаемых радиоактивным источником за определенный промежуток времени; числа требований на выплату страховых сумм за год; числа вызовов, поступающих на телефонную станцию за определенное время суток. Описываемые пуассоновским распределением события, происходящие на постоянной площади или в постоянном объеме, включают: число дефектов на одинаковых образцах вещества; количество бактерий на предметном стекле нескольких микроскопов; число авиационных бомб, упавших на одинаковые площади Лондона в годы второй мировой войны.

Закон Пуассона можно применять для совокупностей, достаточно больших по объему ($n \ge 100$) и имеющих достаточно малую долю единиц, обладающих данным признаком ($p \le 0{,}1$). Данное распределение является предельным для биномиального распределения, если одновременно устремлять число опытов *n* к бесконечности, а вероятность *p* – к нулю, причем их произведение *np* сохраняет постоянное значение:

$$np = \lambda.$$

Это предельное свойство биномиального закона часто находит применение на практике. Допустим, что производится большое количество независимых опытов *n*, в каждом из которых событие *A* имеет очень малую вероятность *p*. Тогда для вычисления вероятности $P_{x,n}$ того, что событие *A* появится ровно *x* раз, можно использовать приближенную формулу

$$P_{x,n} \approx \frac{(np)^x}{x!} e^{-np},$$

где $np = \lambda$ – параметр того закона Пуассона, которым приближенно заменяется биномиальное распределение.

От этого свойства закона Пуассона выражать биномиальное распределение при большом числе опытов и малой вероятности события происходит его название, применяемое в учебниках статистики: *закон редких явлений*.

Рассмотрим один из типичных примеров применения распределения Пуассона для решения производственных задач.

Пример 6.25. На ткацком станке нить обрывается в среднем 1 раз за 4 часа работы станка. Требуется найти вероятность того, что за смену (8 часов) число обрывов нити будет заключено в границах 2 и 4 (не менее 2 и не более 4 обрывов).

Очевидно, что $\lambda = 1/4 \cdot 8 = 2$ – интенсивность обрывов нити за смену.

Для решения задачи используем функцию ПУАССОН.

Если искомую вероятность рассчитывать через плотность распределения (аргумент *интегральная* = 0), то формула для ее нахождения будет иметь следующий вид:

=ПУАССОН(2;2;0)+ПУАССОН(3;2;0)+ПУАССОН(4;2;0).

Данная формула рассчитает значение 0,541 = 0,271 + 0,180 + + 0,090.

Если искомую вероятность вычислять через интегральную функцию распределения (аргумент *интегральная* = 1), то формула для ее нахождения будет иметь следующий вид:

=ПУАССОН(4;2;1)–ПУАССОН(1;2;1).

Данная формула рассчитает значение 0,541 = 0,947 – 0,406.

Таким образом, вероятность того, что за смену на станке будет не менее 2 и не более 4 обрывов нити, равна 0,541.

Математическое ожидание и дисперсия распределения Пуассона равны интенсивности появления события λ:

$$a(x) = \lambda;$$
$$\sigma^2(x) = \lambda.$$

РАЗДЕЛ II

Методы проверки статистических гипотез

ГЛАВА 7
Двухвыборочный z-тест для средних

7.1. Понятие статистической гипотезы

Под *статистической гипотезой* понимают всякое высказывание о генеральной совокупности (случайной величине), проверяемое по выборке (по результатам наблюдений). Процедуру сопоставления высказанной гипотезы с выборочными данными называют *проверкой статистической гипотезы*.

По прикладному содержанию можно выделить следующие основные виды высказываемых в ходе статистической обработки данных гипотез:

- о типе закона распределения исследуемой случайной величины (*см.* описание функции ХИ2РАСП в подразд. 6.3.7);
- об однородности двух или нескольких обрабатываемых выборок или некоторых характеристик анализируемых совокупностей (*см.* главы 7, 8, 9);
- о числовых значениях исследуемой генеральной совокупности (*см.* описание функции СТЬЮДРАСП в подразд. 6.3.8);
- о типе зависимости между компонентами исследуемого многомерного признака (см. подразд. 14.2);
- о независимости и стационарности обрабатываемого ряда наблюдений.

Проверяемую статистическую гипотезу принято называть *основной (или нулевой)* гипотезой (обозначается H_0), а противоречащую ей гипотезу – *альтернативной* (*или конкурирующей*) гипотезой (обозначается H_1).

Поскольку при проверке статистических гипотез приходится иметь дело со статистическим материалом, то, отвергая или принимая нулевую гипотезу, всегда рискуем совершить ошибку. Ошибку, заключающуюся в том, что нулевая гипотеза отвергается, тогда как она в действительности верна, называют *ошибкой первого рода*. Ошибку, состоящую в том, что нулевая гипотеза не отвергается, тогда как она в действительности неверна, называют *ошибкой второго рода*.

Проверка статистических гипотез осуществляется с помощью различных статистических критериев. В качестве критерия используется некоторая случайная величина, значения которой могут быть вычислены на основе имеющихся данных. В множестве возможных значений критерия выбирается подмножество, называемое *критической областью*. Если вычисленное значение критерия принадлежит критической области, то нулевая гипотеза отвергается. Критическая область выбирается таким образом, чтобы вероятность совершить ошибку первого рода не превосходила некоторого заранее определенного положительного числа α. Это число α называют *уровнем значимости* и говорят: «нулевая гипотеза отвергается на уровне значимости α». В качестве α обычно берут одно из чисел: 0,05; 0,01; 0,001.

Вероятность совершить ошибку второго рода обозначается β. Величина $1 - \beta$ называется *мощностью критерия*; она равна вероятности отвергнуть неверную гипотезу.

Чаще всего множество возможных значений критерия принадлежит некоторому интервалу. Интервалом является и критическая область. Граничные точки критической области называются *критическими точками*. Критические точки выбираются таким образом, чтобы при выбранном уровне значимости α мощность критерия $(1 - \beta)$ была *наибольшей*.

Возможны три вида расположения критической области (в зависимости от вида нулевой и альтернативной гипотез, вида и распределения статистического критерия φ):

1) правосторонняя критическая область, состоящая из интервала $(x^{кр}_{пр,\alpha}, +\infty)$; где точка $x^{кр}_{пр,\alpha}$ определяется из условия

$$P(\varphi > x^{кр}_{пр,\alpha}) = \alpha$$

и называется *правосторонней критической точкой*, отвечающей уровню значимости α;

2) левосторонняя критическая область, состоящая из интервала $(-\infty, x^{кр}_{лев,\alpha})$, где точка $x^{кр}_{лев,\alpha}$ определяется из условия

$$P(\varphi < x^{кр}_{лев,\alpha}) = \alpha$$

и называется *левосторонней критической точкой*, отвечающей уровню значимости α;

3) двусторонняя критическая область, состоящая из следующих двух интервалов: $(-\infty, x^{кр}_{лев,\alpha/2})$ и $(x^{кр}_{пр,\alpha/2} +\infty)$, где точки $x^{кр}_{лев,\alpha/2}$ и $x^{кр}_{пр,\alpha/2}$ определяются из условий

$$P(\varphi < x^{кр}_{лев,\alpha/2}) = \alpha/2 \text{ и } P(\varphi > x^{кр}_{пр,\alpha/2}) = \alpha/2$$

и называются *двусторонними критическими точками*.

Наиболее распространенными являются критерии, в основе которых лежат известные распределения: χ^2, Стьюдента, Фишера (*см.* главу 6). Для этих критериев составлены таблицы, в которых указаны критические точки, соответствующие определенным уровню значимости и числу степеней свободы. Порядок использования данных критериев приведен в описаниях одноименных статистических функций и рассмотрен также в главах, посвященных соответствующим режимам обработки статистической информации.

7.2. Краткие сведения из теории статистики

В примере 6.12 был рассмотрен случай использования критерия χ^2 для проверки гипотезы о принадлежности эмпирического распределения к типу нормальных распределений. Здесь рассматривается критерий проверки гипотезы о равенстве средних (математических ожиданий) двух нормальных распределений с *известными* дисперсиями, который находит важное практическое применение.

Действительно, иногда оказывается, что средний результат $\bar{x}$ одной серии наблюдений отличается от среднего результата $\bar{y}$ другой серии. Возникает вопрос: можно ли это различие объяснить случайной ошибкой экспериментов или это отличие не случайно? Иначе говоря, можно ли считать, что результаты экспериментов представляют собой выборки из двух генеральных совокупностей с одинаковыми средними или средние этих совокупностей не равны? Подобная задача возникает, например, при сравнении качества изделий, изготовленных на разных установках.

Рассмотрим формализованную постановку данной задачи.

Изучаются две нормально распределенные случайные величины: $X = N(a_x, \sigma_x)$ и $Y = N(a_y, \sigma_y)$, числовые значения дисперсий σ_x^2 и σ_y^2 которых известны; числовые значения средних a_x и a_y неизвестны.

Пусть $x_1, x_2, \ldots, x_n$ – результаты независимых, проводимых в одинаковых условиях наблюдений величины X, а $y_1, y_2, \ldots, y_m$ – результаты независимых, проводимых в одинаковых условиях наблюдений величины Y.

При сформулированных условиях требуется проверить гипотезу о равенстве математических ожиданий случайных величин X и Y, т.е. гипотезу

$$H_0 : a_X = a_Y.$$

Если гипотеза $H_0 : a_X = a_Y$ принимается, то говорят, что различие выборочных средних $\bar{x}$ и $\bar{y}$ *статистически незначимо.*

В математической статистике доказывается, что если данная гипотеза выполняется, то величина

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}}$$

имеет нормальный закон распределения с нулевым математическим ожиданием и единичной дисперсией, т. е.

$$z = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{\sigma_X^2}{n} + \frac{\sigma_Y^2}{m}}} = N(0,1).$$

Величину z и используют в качестве критерия при проверке гипотезы $H_0 : a_x = a_y$.

7.3. Справочная информация по технологии работы

Режим работы «Двухвыборочный z-тест для средних» служит для проверки гипотезы о различии между средними (математическими ожиданиями) двух нормальных распределений с *известными* дисперсиями.

В диалоговом окне данного режима (рис. 7.1) задаются следующие параметры:

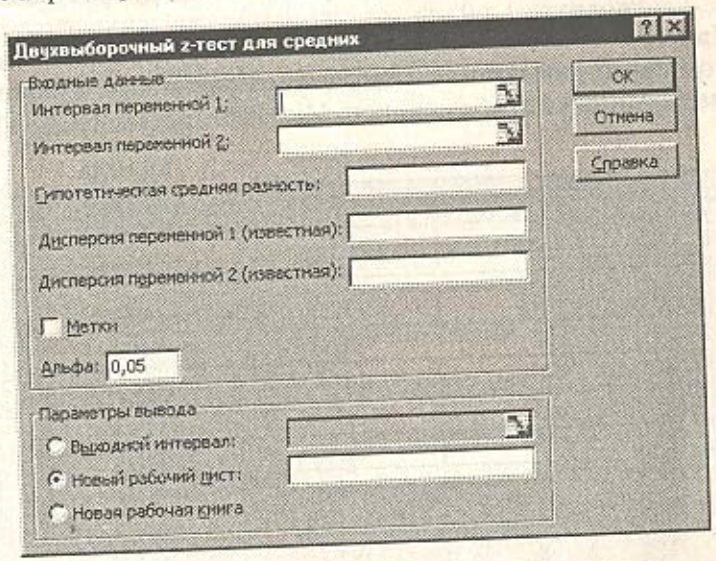

Рис. 7.1

1. *Интервал переменной 1* – вводится ссылка на ячейки, содержащие результаты наблюдений величины X. Диапазон данных должен состоять из одного столбца или одной строки.

2. *Интервал переменной 2* – вводится ссылка на ячейки, содержащие результаты наблюдений величины Y. Диапазон данных должен состоять из одного столбца или одной строки.

3. *Гипотетическая средняя разность* – вводится число, равное предполагаемой разности средних (математических ожиданий) изучаемых генеральных совокупностей. Значение 0 указывает на то, что проверяется гипотеза $H_0 : a_x = a_y$.

4. *Дисперсия переменной 1 (известная)* – вводится *известное* значение дисперсии генеральной совокупности величины *X*.

5. *Дисперсия переменной 2 (известная)* – вводится *известное* значение дисперсии генеральной совокупности величины *Y*.

6. *Метки* – *см.* подразд. 1.1.2.

7. *Альфа* – вводится уровень значимости α, равный вероятности возникновения ошибки первого рода (отвержение нулевой гипотезы).

8. *Выходной интервал/Новый рабочий лист/Новая рабочая книга* – *см.* подразд. 1.1.2.

Пример 7.1. Выборочные данные о диаметре валиков (мм), изготовленных автоматом 1 и автоматом 2, приведены в таблице, сформированной на рабочем листе Microcoft Excel (табл. 7.1) [5].

Таблица 7.1

	C	D	E
23	*N* п/п	Автомат 1	Автомат 2
24	1	182,3	185,3
25	2	183,0	185,6
26	3	181,8	184,8
27	4	181,4	186,2
28	5	181,8	185,8
29	6	181,6	184,0
30	7	183,2	184,2
31	8	182,4	185,2
32	9	182,5	184,2
33	10	179,7	
34	11	179,9	
35	12	181,9	
36	13	182,8	
37	14	183,4	
38	Среднее	182,0	185,0

По выборке объема $n = 14$ найден средний размер $\bar{x} = 182{,}0$ мм диаметра валиков, изготовленных автоматом 1 (ячейка D38 содержит формулу =СРЗНАЧ(D24:D37)). По выборке объема $m = 9$ найден средний размер $\bar{y} = 185{,}0$ мм диаметра валиков, изготовленных автоматом 2 (ячейка E38 содержит формулу =СРЗНАЧ(E24:E32)).

Кроме того, предварительным анализом установлено, что размер диаметра валиков, изготовленных каждым автоматом, имеет нормальный закон распределения с дисперсией $\sigma_x^2 = 5$ мм^2 для автомата 1 и $\sigma_y^2 = 7$мм^2 для автомата 2. Можно ли при уровне значимости $\alpha = 0{,}05$ объяснить различие выборочных средних случайной величиной? Или, иными словами, при уровне значимости $\alpha = 0{,}05$ требуется проверить гипотезу $H_0 : a_x = a_y$.

Для решения задачи используем режим работы «Двухвыборочный *z*-тест для средних». Значения параметров, установленных в одноименном диалоговом окне, представлены на рис. 7.2, а рассчитанные в данном режиме показатели – в табл. 7.2.

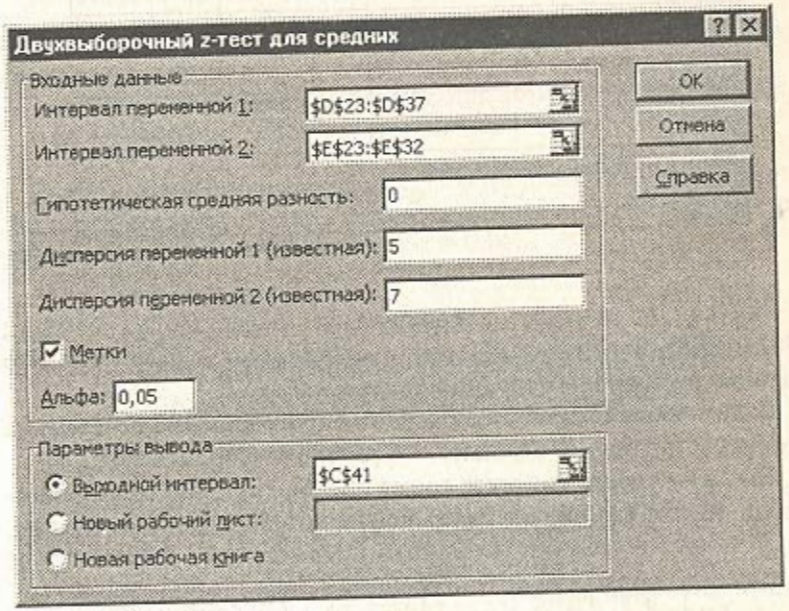

Рис. 7.2

Таблица 7.2

	C	D	E
41	Двухвыборочный z-тест для средних		
42			
43		Автомат 1	Автомат 2
44	Среднее	181,98	185,03
45	Известная дисперсия	5	7
46	Наблюдения	14	9
47	Гипотетическая разность средних	0	
48	z	– 2,867	
49	$P(Z \le z)$ односторонняя	0,002	
50	z критическое одностороннее	1,645	
51	$P(Z \le z)$ двусторонняя	0,004	
52	z критическое двустороннее	1,960	

Так как z_p попадает в критическую область ($|z_p| > |z_{кр}|$; 2,867 > 1,96), то гипотеза $H_0 : a_X = a_Y$ отвергается, т. е. считаем, что различие выборочных средних *неслучайно*.

Дадим более подробное пояснение проведенным расчетам, на основании которых и строился сформулированный вывод.

Так как нулевая гипотеза имеет вид $H_0 : a_X = a_Y$, то альтернативная ей гипотеза будет иметь соответственно вид $H_1 : a_X = a_Y$, т. е. включать в себя два условия: $a_X < a_Y$ и $a_X > a_Y$. В этом случае критическая область будет определяться двумя интервалами $(-\infty, z^{кр}_{лев,\,\alpha/2})$ и $(z^{кр}_{пр,\,\alpha/2}, +\infty)$, где критические точки $z^{кр}_{лев,\,\alpha/2}$ и $z^{кр}_{пр,\,\alpha/2}$ определяются из условий

$$P(z < z^{кр}_{лев,\,\alpha/2}) = \alpha/2 \text{ и } P(z > z^{кр}_{пр,\,\alpha/2}) = \alpha/2,$$

которые с учетом равенства $z_{кр} = N(0,1)$ запишем в следующем виде:

$$P(N(0,1) < z^{кр}_{лев,\,\alpha/2}) = \alpha/2 \text{ и } P(N(0,1) > z^{кр}_{пр,\,\alpha/2}) = \alpha/2.$$

По данной схеме находятся критические точки $z^{кр}_{лев,\,\alpha/2} = -1,96$ и $z^{кр}_{пр,\,\alpha/2} = 1,96$ (показатель *z критическое двустороннее* в табл. 7.2), задающие критическую область $(-\infty; -1,96) \cup (1,96; +\infty)$. Модуль значений критических точек рассчитывается по формуле =НОРМСТОБР(1–0,05/2) в ячейке D52.

Расчетное значение критерия z_p вычисляется в ячейке D48 по формуле

=(D44–E44)/КОРЕНЬ(D45/D46+E45/E46),

где в ячейках D44 и E44 рассчитываются средние значения выборок с помощью функции СРЗНАЧ (*см.* подразд. 4.3);
в ячейках D45 и E45 содержатся значения дисперсий, установленные в диалоговом окне **Двухвыборочный z-тест для средних** (*см.* рис. 7.2);
в ячейках D46 и E46 рассчитываются объемы выборок с помощью функции СЧЕТ (*см.* подразд. 4.3).

Расчетное значение критерия $z_p = -2,867$ попадает в критический интервал $(-\infty; -1,96)$, поэтому нулевая гипотеза $H_0 : a_X = a_Y$ отвергается на уровне значимости $\alpha = 0,05$.

7.4. Статистические функции, связанные с режимом «Двухвыборочный z-тест для средних»

В подразд. 7.3 упоминался ряд статистических функций (СРЗНАЧ, СЧЕТ, НОРМСТОБР), используемых для расчетов в режиме «Двухвыборочный z-тест для средних». Описание этих функций можно найти в подразд. 4.3 и 6.3. Здесь приводится описание функции ZТЕСТ, родственной по своей сущности режиму «Двухвыборочный z-тест для средних».

Функция ZTECT

См. также НОРМРАСП, ДОВЕРИТ.

Синтаксис:

ZTECT (массив; *x*; сигма)

Результат:

Рассчитывает для определенного выборочного массива данных *двустороннее P-значение z-теста*.

Аргументы:

- *массив*: массив данных, с которыми сравнивается *x*;
- *x*: проверяемое значение;
- *сигма*: известное стандартное отклонение генеральной совокупности. Если этот аргумент опущен, то используется оценка генерального стандартного отклонения по выборке.

Замечания:

- если массив пуст, то функция ZTECT помещает в ячейку значение ошибки #Н/Д.

Математико-статистическая интерпретация:

См. описание функций НОРМРАСП, ДОВЕРИТ.

Функция ZTECT служит для проверки гипотезы о числовом значении средней (математического ожидания) нормального распределения при известной дисперсии.

Примечание. Если числовое значение стандартного отклонения генеральной совокупности не известно, то в функции используется оценка стандартного отклонения по представленной выборке (*см.* описание функции СТАНДОТКЛОН в подразд. 4.3).

Рассматривается случайная величина $X = N(a, \sigma)$, причем числовое значение математического ожидания a не известно, а числовое значение дисперсии σ^2 известно.

Выдвигается гипотеза H_0 о том, что среднее (математическое ожидание) равно числу a_0, т. е. $H_0 : a = a_0$. В этом случае альтернативная гипотеза будет иметь вид $H_1 : a \neq a_0$.

В качестве критерия проверки гипотезы берется величина

$$z = \frac{x - a_0}{\sigma / \sqrt{n}},$$

которая при выполнении гипотезы подчиняется нормальному закону распределения с нулевым математическим ожиданием и единичной дисперсией, т. е.

$$z = \frac{\overline{x} - a_0}{\sigma / \sqrt{n}} = N(0,1).$$

Пример 7.2. Результаты девяти выборочных замеров времени изготовления детали (мин) приведены в табл. 7.3, сформированной на рабочем листе Microsoft Excel.

Таблица 7.3

	F	G
23	Номер замера	Время изготовления, мин
24	1	44
25	2	48
26	3	50
27	4	46
28	5	50
29	6	46
30	7	47
31	8	51
32	9	50

Предполагается, что время изготовления – нормально распределенная случайная величина. На уровне значимости $\alpha = 0{,}05$ требуется решить:

1) можно ли принять 50 мин в качестве нормативного времени (математического ожидания) изготовления детали?

2) можно ли принять за норматив 49 мин?

Для варианта 1 проверяется статистическая гипотеза $H_0 : a_x = 50$ мин, а для варианта 2 – гипотеза $H_0 : a_x = 49$ мин.

Расчетные показатели для проверки выдвинутых гипотез приведены в табл. 7.4.

Таблица 7.4

	F	G
23	Номер замера	Время изготовления, мин
24	1	44
25	2	48
26	3	50
27	4	46
28	5	50
29	6	46
30	7	47
31	8	51
32	9	50
33	Среднее	48
34	Оценка стандартного отклонения	2,40
35		
36	z критические двусторонние	−1,96
37	z расчетное ($a_x = 50$)	−2,50
38	z расчетное ($a_x = 49$)	−1,25
39		
40	z расчетное ($a_x = 50$) с помощью функции ZTECT	−2,50
41	z расчетное ($a_x = 49$) с помощью функции ZTECT	−1,25

Содержимое ячеек в табл. 7.4:

• массив G24:G32 содержит исходные данные задачи;

• ячейка G33 содержит формулу =СРЗНАЧ(G24:G32) – рассчитывается среднее значение выборки;

• ячейка G34 содержит формулу =СТАНДОТКЛОН(G24:G32) – оценивается стандартное отклонение по выборке;

• ячейка G36 содержит формулу =НОРМСТОБР(0,05/2) – вычисляются критические точки и тем самым задается критическая область $(-\infty; -1{,}96) \cup (1{,}96; +\infty)$;

• ячейки G37 и G38 содержат соответственно формулы =(G33–50)/G34*3 и =(G33–49)/G34*3, которые вычисляют расчетные значения z-критерия для гипотез $H_0 : a_X = 50$ мин и $H_0 : a_X = 49$ мин (здесь $3 = \sqrt{n} = \sqrt{9}$, n=9 – объем выборки);

• ячейки G40 и G41 содержат соответственно формулы =НОРМСТОБР(1–ZTECT(G24:G32;50)) и =НОРМСТОБР(1–ZTECT(G24:G32;50)), которые вычисляют расчетные значения z-критерия с использованием функции ZTECT, рассчитывающей вероятностные значения z-теста.

Примечание. В постановке задачи не приведена информация о значении генерального стандартного отклонения, поэтому при решении использовалась оценка генерального стандартного отклонения по представленной выборке.

При проверке гипотезы $H_0 : a_X = 50$ мин расчетное значение критерия $z_p = -2{,}50$ попадает в критический интервал $(-\infty; -1{,}96)$, поэтому данная гипотеза отвергается, а принимается альтернативная гипотеза $H_1 : a_X = 48$ мин (среднее значение, вычисленное по представленной выборке). Или, иначе говоря, 50 мин нельзя считать нормативным временем изготовления детали и за норматив берем 48 мин.

При проверке гипотезы $H_0 : a_X = 49$ мин расчетное значение критерия $z_p = -1{,}25$ не попадает в критическую область $(-\infty; -1{,}96) \cup (1{,}96; +\infty)$, поэтому данная гипотеза не отвергается, т.е. за норматив времени изготовления детали берем 49 мин.

Заметим, что функция ZTECT аналогична функции НОРМРАСП при условии, что в качестве аргумента σ используется аргумент μ, выражающий стандартное отклонение выборочной средней от генеральной средней и получивший название *средней*

ошибки выборки (*см.* описание функции ДОВЕРИТ в подразд. 6.3.1).

Зная, что

$$\mu = \frac{\sigma}{\sqrt{n}},$$

можно вывести формулу

$$\text{ZTECT}(\textit{массив}, x) = \text{НОРМРАСП}\left(\frac{\bar{x} - x}{\mu}\right) =$$

$$= \text{НОРМРАСП}\left(\frac{\sqrt{n}(\bar{x} - x)}{\sigma}\right).$$

Пример 7.3. В массив H3:H12 введены следующие значения случайной величины X {4; 3; 5; 8; 12; 8; 6; 10; 3; 5}. Тогда функция =ZTECT(H3:H12;7) рассчитает значение 0,735. Это же значение вычислит и функция =НОРМРАСП(7;H14;H15/H16;1), если:

- ячейка H14 содержит формулу =СРЗНАЧ(H3:H12), которая определяет значение выборочной средней;
- ячейка H15 содержит формулу =СТАНДОТКЛОН(H3:H12), которая оценивает значение генерального стандартного отклонения по представленной выборке;
- ячейка H16 содержит формулу =КОРЕНЬ(10), которая рассчитывает значение квадратного корня из размера выборочной совокупности, т.е. $\sqrt{n}$.

ГЛАВА 8
Двухвыборочный *t*-тест с одинаковыми и различными дисперсиями

8.1. Краткие сведения из теории статистики

В главе 7 была рассмотрена процедура проверки гипотезы о равенстве средних (математических ожиданий) двух нормальных распределений с *известными* дисперсиями.

Настоящая глава посвящена процедурам проверки гипотез о равенстве средних (математических ожиданий) двух нормальных распределений с *неизвестными* дисперсиями. Причем относительно параметров σ_X^2 и σ_Y^2 можно выдвинуть следующие два предположения:

1) обе дисперсии неизвестны, но предполагается, что они равны между собой ($\sigma_X^2 = \sigma_Y^2$);

2) обе дисперсии неизвестны, их равенство не предполагается ($\sigma_X^2 \neq \sigma_Y^2$).

В случае когда обе дисперсии неизвестны, но предполагаются равными между собой, имеем дело с двумя оценками $s_X^2 = s_Y^2$ одной и той же величины дисперсии $\sigma_X^2 = \sigma_Y^2$. В связи с этим разумно перейти к объединенной оценке σ^2:

$$s^2 = \frac{s_X^2(n-1) + s_Y^2(m-1)}{(n-1)+(m-1)}.$$

В математической статистике доказывается, что если гипотеза $H_0: a_X = a_Y$ выполняется, то величина

$$t = \frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n} + \frac{1}{m}}}$$

имеет распределение Стьюдента с $k = n+m-2$ степенями свободы, т. е.

$$\frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n} + \frac{1}{m}}} = t(k = n+m-2).$$

Величину t и используют в качестве критерия при проверке гипотезы $H_0: a_X = a_Y$.

Когда дисперсии неизвестны и их равенство не предполагается ($\sigma_X^2 \neq \sigma_Y^2$), используется аналог z-статистики (*см.* подразд. 7.2) с заменой неизвестных дисперсий их оценками

$$t = \frac{\bar{x} - \bar{y}}{\sqrt{\frac{s_X^2}{n} + \frac{s_Y^2}{m}}}.$$

В этой ситуации указать точное распределение введенной статистики затруднительно. Известно, однако, что это распределение близко к распределению Стьюдента с числом степеней свободы, равным

$$k = \frac{(s_X^2/n + s_Y^2/m)^2}{\frac{(s_X^2/n)^2}{n-1} + \frac{(s_Y^2/m)^2}{m-1}}.$$

Последний статистический критерий (при $\sigma_X^2 \neq \sigma_Y^2$) называют также *критерием Фишера–Беренса*. Данный критерий не часто применяют на практике, потому что, даже когда неизвестные дисперсии σ_X^2 и σ_Y^2 существенно различны, предположение, что они на самом деле равны, дает результаты, довольно близкие к получаемым по этому критерию, но при гораздо меньшем объеме вычислений.

Примечание. Строго говоря, описанные выше критерии применимы только к выборкам, извлеченным из *нормальной* генеральной совокупности. Вместе с тем специальные исследования показали, что t-критерий является (особенно при больших объемах выборок n) весьма устойчивым по отношению к отклонениям исследуемых генеральных совокупностей от нормальных. А это значит, что он может применяться и к выборкам из негауссовских генеральных совокупностей с той лишь оговоркой, что истинные значения уровня значимости и мощности критерия в этом случае будут незначительно отличаться от заданных.

8.2. Справочная информация по технологии работы

Режимы работы «Двухвыборочный t-тест с одинаковыми дисперсиями» (*гомоскедастический тест*) и «Двухвыборочный t-тест с различными дисперсиями» (*гетероскедастический тест*) служат

для проверки гипотез о различии между средними (математическими ожиданиями) двух нормальных распределений соответственно с неизвестными, но равными дисперсиями ($\sigma_X^2 = \sigma_Y^2$) и с неизвестными дисперсиями, равенство которых не предполагается ($\sigma_X^2 \neq \sigma_Y^2$).

В диалоговых окнах данных режимов (рис. 8.1 и 8.2) задаются параметры, аналогичные параметрам, задаваемым в диалоговом окне **Двухвыборочный *z*-тест для средних** (*см.* рис. 7.1), только отсутствуют поля *Дисперсия переменной 1* (*известная*) и *Дисперсия переменной 2* (*известная*).

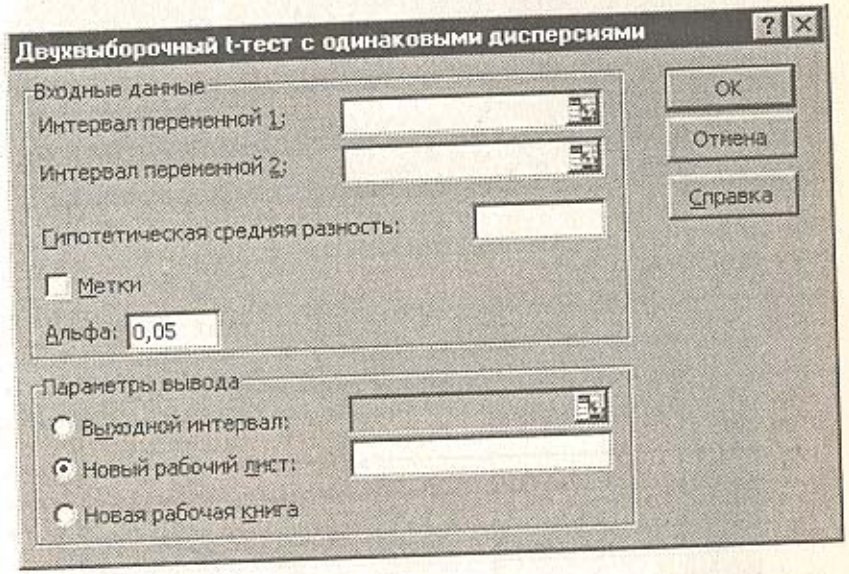

Рис. 8.1

Пример 8.1. Выборочные данные о расходе сырья при производстве продукции по старой и новой технологиям приведены в табл. 8.1, сформированной на рабочем листе Microsoft Excel [5].

При уровне значимости $\alpha = 0{,}05$ требуется проверить гипотезу $H_0 : a_X = a_Y$, предположив, что соответствующие генеральные совокупности X и Y имеют нормальные распределения:

1) с одинаковыми дисперсиями σ_X^2 и σ_Y^2;

2) с различными дисперсиями σ_X^2 и σ_Y^2.

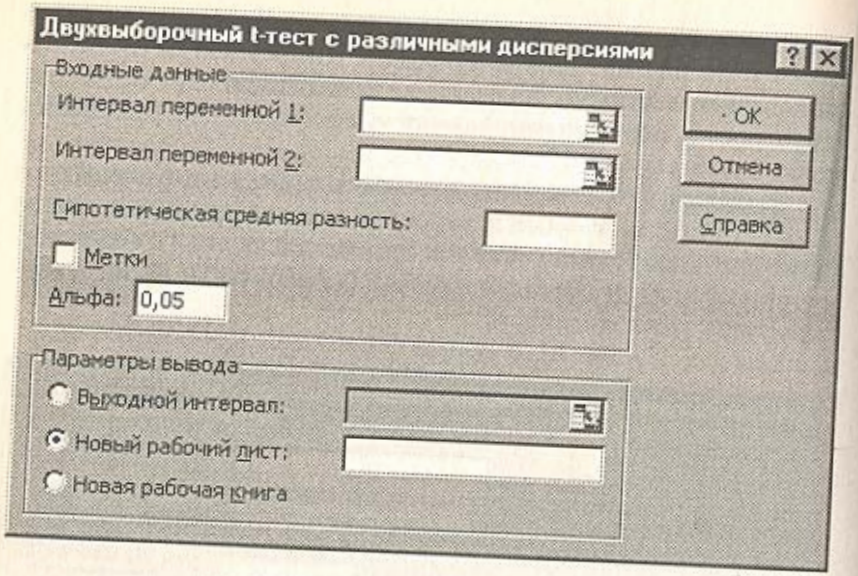

Рис. 8.2

Таблица 8.1

	C	D	E
17	Номер изделия	Старая технология	Новая технология
18	1	308	308
19	2	308	304
20	3	307	306
21	4	308	306
22	5	304	306
23	6	307	304
24	7	307	304
25	8	308	304
26	9	307	306
27	10		304
28	11		303
29	12		304
30	13		303

Для проверки предположения 1 используем режим работы «Двухвыборочный *t*-тест с одинаковыми дисперсиями», а для проверки предположения 2 – «Двухвыборочный *t*-тест с различными дисперсиями». Значения параметров, установленных в одноименных диалоговых окнах, представлены на рис. 8.3 и 8.4, а рассчитанные в этих режимах показатели – в табл. 8.2 и 8.3 соответственно.

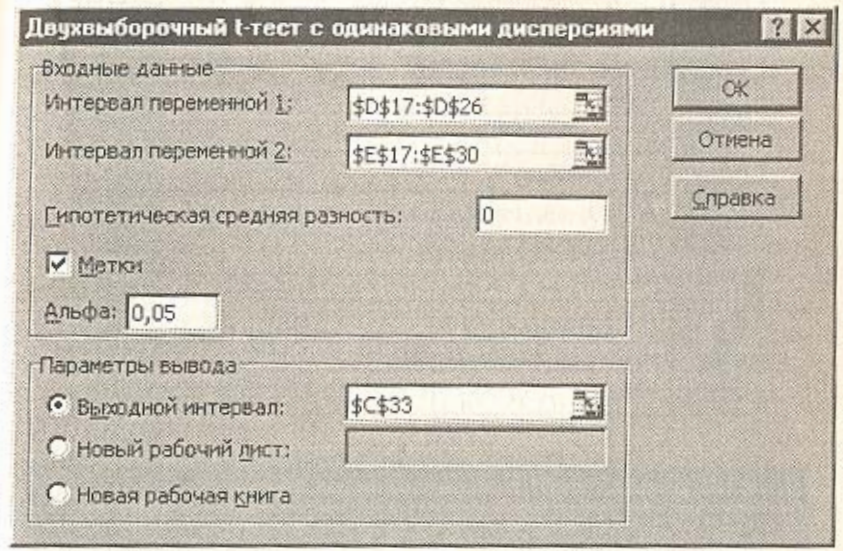

Рис. 8.3

Таблица 8.2

	C	D	E
33	**Двухвыборочный t-тест с одинаковыми дисперсиями**		
34			
35		*Старая технология*	*Новая технология*
36	Среднее	307,11	304,77
37	Дисперсия	1,61	2,19
38	Наблюдения	9	13
39	Объединенная дисперсия	1,96	

Продолжение

	C	D	E
33	**Двухвыборочный t-тест с одинаковыми дисперсиями**		
34			
35		*Старая технология*	*Новая технология*
40	Гипотетическая разность средних	0	
41	*df*	20	
42	*t*-статистика	3,86	
43	$P(T \le t)$ односторонняя	0,0005	
44	*t* критическое одностороннее	1,72	
45	$P(T \le t)$ двусторонняя	0,0010	
46	*t* критическое двустороннее	2,09	

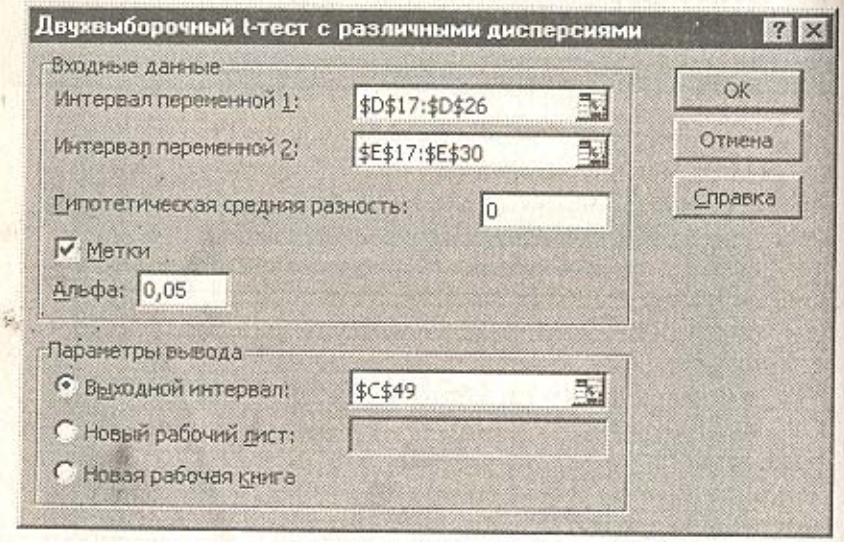

Рис. 8.4

Таблица 8.3

	C	D	E
49	**Двухвыборочный t-тест с различными дисперсиями**		
50			
51		*Старая технология*	*Новая технология*
52	Среднее	307,11	304,77
53	Дисперсия	1,61	2,19
54	Наблюдения	9	13
55	Гипотетическая разность средних	0	
56	*df*	19	
57	*t*-статистика	3,97	
58	$P(T \le t)$ односторонняя	0,0004	
59	*t* критическое одностороннее	1,73	
60	$P(T \le t)$ двусторонняя	0,0008	
61	*t* критическое двустороннее	2,09	

Для предположения 1 $t_p = 3{,}86$, а критическая область образуется интервалами $(-\infty; -2{,}09) \cup (2{,}09; +\infty)$. Для предположения 2 $t_p = 3{,}97$, а критическая область образуется интервалами $(-\infty; -2{,}09) \cup (2{,}09; +\infty)$. Так как t_p в обоих случаях попадает в критический интервал $(2{,}09; +\infty)$, то гипотезу $H_0 : a_X = a_Y$ отвергаем, т. е. при переходе на новую технологию происходит изменение среднего расхода сырья на одно изделие. При этом, конечно, следует иметь в виду, что данное заключение может оказаться ошибочным (на самом деле $a_X = a_Y$), т. е. имеет место ошибка первого рода, вероятность которой равна $\alpha = 0{,}05$.

Заметим, что и в первом, и во втором случае получены результаты, несущественно отличающиеся друг от друга (в первом случае $t_p' = 3{,}86$, во втором случае $t_p = 3{,}97$). Данное обстоятельство

еще раз подтверждает, что для проверки гипотезы $H_0 : a_X = a_Y$ при предположении $\sigma_X^2 \neq \sigma_Y^2$ можно пользоваться и критерием

$$\frac{\bar{x} - \bar{y}}{s\sqrt{\frac{1}{n} + \frac{1}{m}}} = t(k = n + m - 2),$$

особенно в тех случаях, когда предполагается, что σ_X^2 и σ_Y^2 различаются незначительно.

Рассмотрим более подробно механизм расчетов основных показателей, представленных в результирующих табл. 8.2 и 8.3.

В первом случае (*см.* табл. 8.2) расчетное значение критерия t_p вычисляется в ячейке D42 по формуле

```
=(D36–E36)/КОРЕНЬ(D39)/КОРЕНЬ(1/D38+1/E38):
```

где в ячейках D36 и E36 рассчитываются средние значения выборок с помощью функции СРЗНАЧ (*см.* подразд. 4.3);
в ячейках D38 и E38 определяются объемы выборок с помощью функции СЧЕТ (см. подразд. 4.3);
в ячейке D39 вычисляется оценка объединенной дисперсии, рассчитываемая, в свою очередь, по формуле

```
=(D37*(D38–1)+E37*(E38–1))/((D38–1)+(E38–1)),
```

где в ячейках D37 и E37 вычисляются оценки дисперсий с помощью функции ДИСП (*см.* подразд. 4.3).

Число степеней свободы (показатель *df*) рассчитывается в ячейке D41 по формуле =D38+E38–2, а модуль значения критических точек (показатель *t критическое двустороннее*) вычисляется в ячейке D46 по формуле =СТЬЮДРАСПОБР(0,05;D41).

Во втором случае (*см.* табл. 8.3) расчетное значение критерия t_p вычисляется в ячейке D57 по формуле

```
=(D52–E52)/КОРЕНЬ(D53/D54+E53/E54),
```

где в ячейках D52 и E52 рассчитываются средние значения выборок с помощью функции СРЗНАЧ;

в ячейках D53 и E53 вычисляются оценки дисперсий с помощью функции ДИСП;
в ячейках D54 и E54 определяются объемы выборок с помощью функции СЧЕТ.

Число степеней свободы (показатель *df*) рассчитывается в ячейке D56 по формуле

```
=((D53/D54+E53/E54)^2)/((D53/D54)^2/(D54–1)+
+(E53/E54)^2/(E54–1)),
```

после чего оно округляется до целого числа с помощью функции ОКРУГЛ (здесь $k = 18{,}96$, после округления которого показатель $df = 19$).

Модуль значения критических точек (показатель *t критическое двустороннее*) рассчитывается в ячейке D61 по формуле

```
=СТЬЮДРАСПОБР(0,05;D56).
```

8.3. Статистические функции, связанные с режимами «Двухвыборочный *t*-тест с одинаковыми дисперсиями» и «Двухвыборочный *t*-тест с различными дисперсиями»

В подразд 8.2 упоминался ряд статистических функций (СРЗНАЧ, СЧЕТ, ДИСП, СТЬЮДРАСПОБР), используемых для производства расчетов в режимах «Двухвыборочный *t*-тест с одинаковыми дисперсиями» и «Двухвыборочный *t*-тест с различными дисперсиями». Описание этих функций можно найти в подразд. 4.3.

Здесь приводится описание функции ТТЕСТ, родственной по своей сущности упомянутым режимам.

Функция ТТЕСТ

См. также СТЬЮДРАСП, СТЬЮДРАСПОБР.

Синтаксис:

ТТЕСТ (массив1; массив2; хвосты; тип)

Результат:

Рассчитывает для двух выборочных массивов данных *одностороннее* или *двустороннее* *P*-значение *t*-теста.

Аргументы:

- *массив1:* первое множество выборочных данных;
- *массив2*: второе множество выборочных данных;
- *хвосты:* число хвостов распределения. Если аргумент *хвосты* = 1, то функция ТТЕСТ использует одностороннее распределение, если аргумент *хвосты* = 2 – двустороннее распределение;
- *тип:* вид исполняемого *t*-теста.

Значение аргумента *тип*	Вид выполняемого теста
1	Двухвыборочный парный*
2	Двухвыборочный с неравными дисперсиями (гетероскедастический)
3	Двухвыборочный с равными дисперсиями (гомоскедастический)

Замечания:

- если аргументы *массив1* и *массив2* имеют различное число точек данных, а аргумент *тип* = 1 (парный), то функция ТТЕСТ помещает в ячейку значение ошибки #Н/Д;
- аргументы *хвосты* и *тип* усекаются до целых;
- если аргумент *хвосты* или *тип* не являются числом, то функция ТТЕСТ помещает в ячейку значение ошибки #ЗНАЧ!;
- если аргумент *хвосты* имеет значение, отличное от 1 и 2, функция ТТЕСТ помещает в ячейку значение ошибки #ЧИСЛО!.

Математико-статистическая интерпретация:

См. подразд. 7.1, 8.1, а также описание функций СТЬЮДРАСП и СТЬЮДРАСПОБР (подразд. 6.3.8).

♦ В примере 8.1 функция ТТЕСТ может использоваться для расчета одностороннего (ячейка D43 в табл. 8.2, ячейка D58 в табл. 8.3) и двустороннего (ячейка D45 в табл. 8.2, ячейка D60 в табл. 8.3) *P*-значения *t*-теста. Данное *P*-значение определяет уровень значимости, соответствующий расчетному критерию t_p, и вычисляется в рассмотренных режимах с помощью функции СТЬЮДРАСП.

*См. главу 10.

В первом случае (*см.* табл. 8.2) одностороннее *P*-значение *t*-теста рассчитывается по формуле

=СТЬЮДРАСП(D42;D41;1),

которая адекватна формуле

=ТТЕСТ(D18:D26;E18:E30;1;2),

а двустороннее *P*-значение *t*-теста вычисляется по формуле

=СТЬЮДРАСП(D42;D41;2),

которая адекватна формуле

=ТТЕСТ(D18:D26;E18:E30;2;2).

Во втором случае (*см.* табл. 8.3) *P*-значения, рассчитываемые функцией ТТЕСТ, несколько отличаются от *P*-значений, вычисляемых функцией СТЬЮДРАСП, используемой в режиме «Двухвыборочный *t*-тест с различными дисперсиями». Это объясняется тем, что функция ТТЕСТ при расчете *P*-значений учитывает возможный дробный характер числа степеней свободы, в то время как функция СТЬЮДРАСП усекает число степеней свободы до целого числа (*см.* описание функции СТЬЮДРАСП в подразд. 6.3.8).

В табл. 8.3 одностороннее *P*-значение *t*-теста, равное 0,0004085, рассчитывается по формуле =СТЬЮДРАСП(D57;D56; 1) при $k = 19$. Сравните со значением 0,0004101, вычисляемым по формуле =ТТЕСТ(D18:D26;E18:E30;1;3), где $k = 18{,}96$. Аналогично двустороннее *P*-значение *t*-теста, равное 0,0008171, рассчитывается по формуле =СТЬЮДРАСП(D57;D56;2). Сравните со значением 0,0008202, вычисляемым по формуле =ТТЕСТ(D18:D26; E18:E30;2;3).

ГЛАВА 9
Двухвыборочный *F*-тест для дисперсий

9.1. Краткие сведения из теории статистики

В главе 8 были рассмотрены процедуры проверки гипотез о равенстве средних (математических ожиданий) двух нормальных

распределений с *неизвестными* дисперсиями. При этом относительно параметров σ_X^2 и σ_Y^2 выдвигались два возможных предположения: 1) $\sigma_X^2 = \sigma_Y^2$ и 2) $\sigma_X^2 \neq \sigma_Y^2$. Как, не располагая всеми сведениями о генеральных совокупностях, а имея лишь выборки из них, убедиться, например, в приемлемости гипотезы о равенстве генеральных дисперсий?

Настоящая глава и посвящена решению этого вопроса, т.е. в ней рассматривается процедура проверки гипотезы *о равенстве* дисперсий двух нормальных распределений.

Отметим, что эта задача имеет и самостоятельное значение. Дисперсия характеризует точность работы приборов, технологических процессов и т.д.; убедившись в равенстве двух дисперсий, мы тем самым убеждаемся, например, в том, что два прибора, два технологических процесса обеспечивают одинаковую точность.

В математической статистике доказывается, что если гипотеза $H_0 : \sigma_X^2 = \sigma_Y^2$ выполняется, то величина

$$F = \frac{s_X^2}{s_Y^2}$$

имеет F-распределение с $k = n-1$ и $l = m-1$ числом степеней свободы, т.е.

$$\frac{s_X^2}{s_Y^2} = F(k = n-1, l = m-1).$$

Величину F, называемую дисперсионным отношением Фишера (или просто статистикой Фишера), и используют в качестве критерия при проверке гипотезы $H_0 : \sigma_X^2 = \sigma_Y^2$.

Поскольку величина F является неотрицательной, критическая область данной величины будет принадлежать интервалу $(0; +\infty)$.

Примечание. F-критерий является чувствительным к нарушениям предположения о нормальности.

9.2. Справочная информация по технологии работы

Режим работы «Двухвыборочный F-тест для дисперсий» служит для проверки гипотезы H_0 о равенстве дисперсий двух нормальных распределений. При этом в качестве альтернативной рассматривается гипотеза $H_1 : \sigma_X^2 < \sigma_Y^2$, если $s_X^2 < s_Y^2$; или гипотеза $H_1 : \sigma_X^2 > \sigma_Y^2$, если $s_X^2 > s_Y^2$.

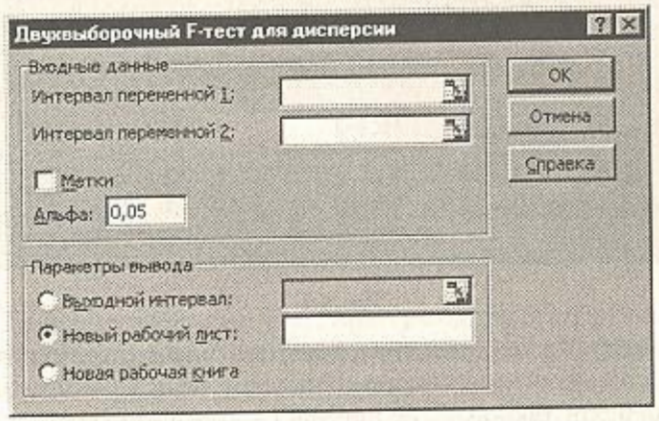

Рис. 9.1

В диалоговом окне данного режима (рис. 9.1) задаются параметры, аналогичные параметрам, задаваемым в диалоговом окне **Двухвыборочный z-тест для средних** (*см.* рис. 7.1), только отсутствуют поля *Дисперсия переменной 1 (известная)*, *Дисперсия переменной 2 (известная)* и *Гипотетическая средняя разность*.

Пример 9.1. Выборочные данные о расходе сырья по старой и новой технологиям приведены в таблице, сформированной на рабочем листе Microsoft Excel [5] (см. табл. 8.1). Можно ли при уровне значимости $\alpha = 0{,}05$ считать статистически незначимым различие между оценками $s_X^2 = 1{,}61$ и $s_Y^2 = 2{,}19$, рассчитанными в табл. 8.2.

Для решения задачи используем режим работы «Двухвыборочный F-тест для дисперсий». Значения параметров, установленных

в одноименном диалоговом окне, представлены на рис. 9.2, а рассчитанные в данном режиме показатели – в табл. 9.1.

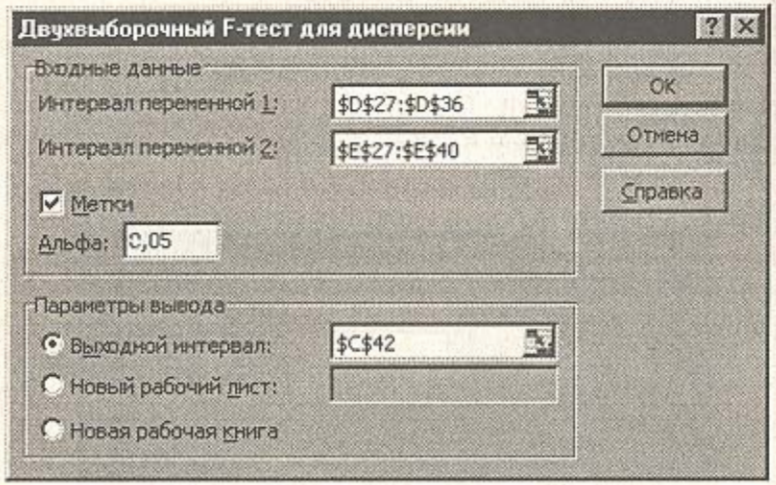

Рис. 9.2

Из табл. 9.1 видно, что расчетное значение F-критерия $F_р = 0{,}73$, а критическая область образуется левосторонним интервалом (0; 0,30). Так как $F_р$ не попадает в критическую область, то гипотезу о равенстве дисперсий расхода сырья при старой и новой технологиях принимаем.

Рассмотрим более подробно механизм расчета основных показателей, представленных в результирующей табл. 9.1.

Расчетное значение критерия $F_р$ вычисляется в ячейке D49 по формуле

=D46/E46,

где в ячейках D36 и E46 рассчитываются оценки дисперсий с помощью функции ДИСП (*см.* описание в подразд. 4.3).

Число степеней свободы (показатель *df*) рассчитывается в ячейках D48 и E48 по формулам =D47–1 и =E47–1 соответственно. Значение левосторонней критической точки $F^{кр}_{лев,\,\alpha}$ (показатель *F критическое одностороннее*) определяется в ячейке D51 по формуле =FРАСПОБР(1–0,05;D48;E48).

Таблица 9.1

	C	D	E
42	Двухвыборочный F-тест для дисперсии		
43			
44		*Старая технология*	*Новая технология*
45	Среднее	307,11	304,77
46	Дисперсия	1,61	2,19
47	Наблюдения	9	13
48	*df*	8	12
49	F	0,73	
50	$P(F \le f)$ односторонняя	0,34	
51	*F* критическое одностороннее	0,30	

В отличие от ранее рассмотренных режимов проверки статистических гипотез (*см.* главы 7, 8), в режиме «Двухвыборочный *F*-тест для дисперсий» рассчитываются только односторонние оценки *P*-значения (ячейка D50) и $F^{кр}_{лев,\,\alpha}$ (ячейка D51). Это объясняется тем, что в данном режиме при проверке гипотезы $H_0 : \sigma^2_X = \sigma^2_Y$ в качестве альтернативной рассматривается гипотеза $H_1 : \sigma^2_X < \sigma^2_Y$ (если $s^2_X < s^2_Y$) или гипотеза $H_1 : \sigma^2_X > \sigma^2_Y$ (если $S^2_X > S^2_Y$).

Чтобы получить двустороннюю оценку для $F_{кр}$ (в этом случае в качестве альтернативной рассматривается гипотеза $H_1 : \sigma^2_X = \sigma^2_Y$), необходимо использовать функцию FРАСПОБР при уровне значимости $\alpha/2 = 0{,}025$. Тогда формула =FРАСПОБР(1–0,025;D48;E48) рассчитает значение левосторонней критической точки $F^{кр}_{лев,\,\alpha/2} = 0{,}24$, а формула =FРАСПОБР(0,025;D48;E48) значение правосторонней критической точки $F^{кр}_{лев,\,\alpha/2} = 3{,}51$. Таким образом, при двусторонней оценке будем иметь критическую область, являющуюся объединением двух интервалов $(0; 0{,}24) \cup (3{,}8; +\infty)$. Но и в этом случае $F_р = 0{,}73$ не принадлежит ни одному из критических интервалов, поэтому гипотеза $H_0 : \sigma^2_X = \sigma^2_Y$ принимается.

9.3. Статистические функции, связанные с режимом «Двухвыборочный *F*-тест для дисперсий»

В подразд. 9.2 упоминался ряд статистических функций (ДИСП, FРАСПОБР), используемых для производства расчетов в режиме «Двухвыборочный *F*-тест для дисперсий». Описание этих функций можно найти в подразд. 4.3 и 6.3.9.

Здесь приводится описание функции ФТЕСТ, родственной по своей сущности режиму «Двухвыборочный *F*-тест для дисперсий».

Функция ФТЕСТ

См. также FРАСП, FРАСПОБР.

Синтаксис:

ФТЕСТ (массив1; массив2)

Результат:

Рассчитывает для двух выборочных массивов данных *двустороннее P*-значение *F*-теста.

Аргументы:

- *массив1*: первое множество выборочных данных;
- *массив2*: второе множество выборочных данных.

Замечания:

- аргументы должны быть числами или именами, массивами или ссылками, содержащими числа;
- если аргумент, который является массивом или ссылкой, содержит текстовые, логические значения или пустые ячейки, то такие значения игнорируются; однако ячейки с нулевыми значениями учитываются;
- если количество точек данных в аргументе *массив1* или *массив2* меньше 2 или если дисперсия аргумента *массив1* или *массив2* равна 0, то функция ФТЕСТ помещает в ячейку значение ошибки #ДЕЛ/0!.

Математико-статистическая интерпретация:

См. подразд. 7.1 и 9.1, а также описание функций FРАСП и FРАСПОБР в подразд. 6.3.9.

♦ В примере 9.1 функция ФТЕСТ может использоваться для расчета одностороннего *P*-значения *F*-теста (ячейка D50 в табл. 9.1). Данное *P*-значение определяет уровень значимости, соответствующий расчетному критерию F_p, и вычисляется в режиме «Двухвыборочный *F*-тест для дисперсий» с помощью функции FРАСП.

Например, в табл. 9.1 одностороннее *P*-значение *F*-теста рассчитывается по формуле

=1-FРАСП(D49;D48;E48),

которая адекватна формуле

=ФТЕСТ(D28:D36;E28:E40)/2.

Примечания: 1. Функция ФТЕСТ рассчитывает двустороннее *P*-значение *F*-теста, поэтому для рассматриваемого случая это значение делится на 2.

2. Если $s_X^2 > s_Y^2$, то *P*-значение рассчитывается с помощью функцией FРАСП (без вычета из 1).

ГЛАВА 10
Парный двухвыборочный *t*-тест для средних

10.1. Краткие сведения из теории статистики

Рассмотренные в главе 8 процедуры сравнения двух выборок часто применяются для обнаружения результата какого-либо воздействия либо, напротив, для подтверждения его отсутствия. Чем более однородными окажутся выбранные для эксперимента объекты (для контроля и воздействия), чем меньше их случайные различия, тем точнее можно будет дать ответ на поставленный вопрос. Ясно, что различие между объектами, выбранными для воздействия и для контроля (или для двух разных воздействий, если интерес представляет их сопоставление), будет наименьшим, если в обоих качествах выступает один и тот же объект. Если это возможно, то далее обычным порядком составляется группа экспериментальных объектов и затем для каждого объекта измеряются два значения интересующей нас характеристики (например, до воздействия и после или при двух разных воздействиях). Так возникают пары наблюдений или парные данные.

Пусть x_i и y_i – результаты измерений для объекта номер i, $i = 1, ..., n$, где n – численность экспериментальной группы (число объектов). Тогда совокупность пар случайных величин $(x_1, y_1), ..., (x_n, y_n)$ образует парные данные.

Как обычно, все наблюдения будем считать реализациями случайных величин и предполагать, что методика эксперимента обеспечивает их независимость для разных объектов. Но наблюдения, входящие в одну пару, нельзя считать независимыми, поскольку они относятся к одному и тому же объекту. Эти два наблюдения отражают свойства общего для них индивидуального объекта и потому могут зависеть друг от друга.

Для пар наблюдений (x_i, y_i) введем величину $z_i = y_i - x_i$, которую будем считать независимой и нормально распределенной. Тем самым задача о парных данных сводится к задаче *об одной нормальной выборке при неизвестной дисперсии*.

При неизвестном числовом значении дисперсии σ_Z^2 в основу проверки гипотезы

$$H_0 : a_Z = a_0,$$

где a_0 – заранее заданное число,

положен критерий

$$t = \frac{\bar{z} - a_0}{s/\sqrt{n}}, \qquad (10.1)$$

который при выполнении гипотезы $H_0 : a_Z = a_0$ имеет t-распределение с числом степеней свободы $k = n - 1$, т. е.

$$\frac{\bar{z} - a_0}{s/\sqrt{n}} = t(k = n-1).$$

10.2. Справочная информация по технологии работы

Режим работы «Парный двухвыборочный t-тест для средних» служит для проверки гипотезы о различии между средними (математическими ожиданиями) двух нормальных распределений на основе *парных* выборочных данных. При этом равенство дисперсий генеральных совокупностей не предполагается ($\sigma_X^2 \neq \sigma_Y^2$).

В диалоговом окне данного режима (рис.10.1) задаются параметры, аналогичные параметрам, задаваемым в диалоговом окне **Двухвыборочный z-тест для средних** (*см.* рис. 7.1), только отсутствуют поля *Дисперсия переменной 1 (известная)* и *Дисперсия переменной 2 (известная)*.

Пример 10.1. Каждый из n образцов проволоки разламывают на два куска, для одного (выбор производится случайно) измеряется нагрузка на растяжение при фиксированной низкой температуре, а для другого – при фиксированной высокой. Результаты измерений приведены в табл. 10.1, сформированной на рабочем листе Microsoft Excel.

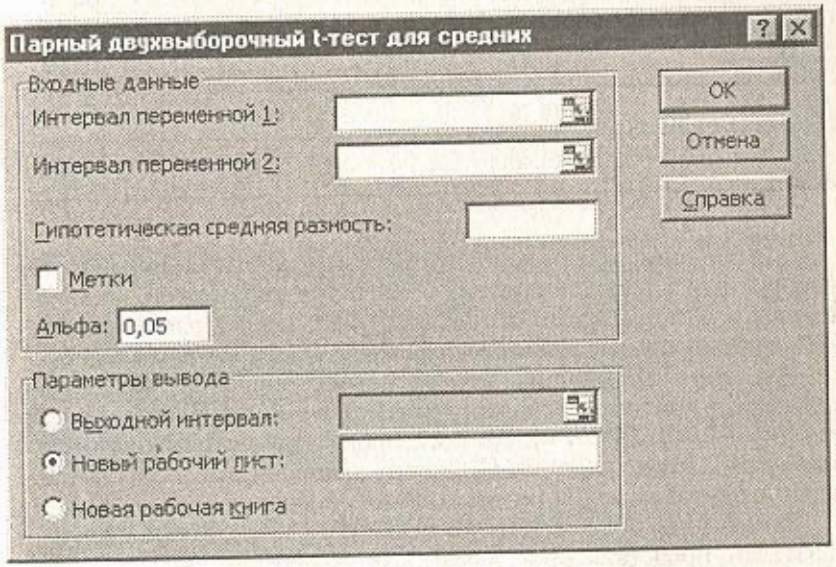

Рис. 10.1

Требуется проверить, влияет ли разность температур на величину растяжения.

Таблица 10.1

	C	D	E	F
65	Номер образца	Измерение при низкой температуре, см	Измерение при высокой температуре, см	Разность, см
66	1	10,40	10,41	–0,01
67	2	10,36	10,38	–0,02
68	3	10,38	10,38	–0,00
69	4	10,41	10,43	–0,02
70	5	10,43	10,44	–0,01
71	6	10,42	10,42	–0,00
72	7	10,39	10,40	–0,01
73	8	10,41	10,42	–0,01
74	9	10,38	10,38	–0,00
75	10	10,40	10,41	–0,01

Для решения задачи используем режим работы «Парный двухвыборочный *t*-тест для средних». Значения параметров, установленных в одноименном диалоговом окне, представлены на рис. 10.2, а рассчитанные в данном режиме показатели – в табл. 10.2.

Из табл. 10.2 видно, что расчетное значения *t*-критерия $t_p = -3{,}86$, а критическая область образуется объединением интервалов $(-\infty; -2{,}26)$ $(2{,}26; +\infty)$. Так как t_p попадает в критический интервал $(-\infty; -2{,}26)$, то гипотезу $H_0 : a_X = a_Y$ отвергаем, т. е. разность температур влияет на величину растяжения проволоки.

Рассмотрим более подробно механизм расчетов основных показателей, представленных в результирующей табл. 10.2.

Расчетное значение критерия t_p вычисляется в ячейке D87 по формуле

=СРЗНАЧ(F66:F75)/СТАНДОТКЛОН(F66:F75)*
*КОРЕНЬ(СЧЕТ(F66:F75)),

которая соответствует математической формуле (10.1),

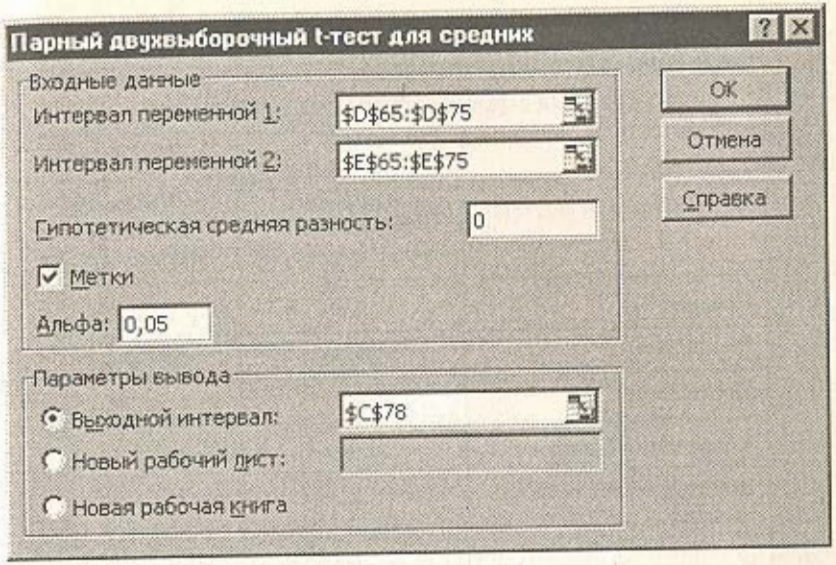

Рис. 10.2

где $a_0 = 0$ (показатель *Гипотетическая разность средних*).

Примечание. На самом деле ячейки столбца *F* (*см.* табл. 10.1) в производстве расчетов не участвуют и приведены здесь как промежуточные результаты для пояснения формул. Значения столбца *F* легко вычисляются на основании данных столбцов *D* и *E* с помощью формулы массива {=D66:D75–E66 :E75}.

Число степеней свободы (показатель *df*) определяется в ячейке D86 по формуле =D83–1, а модуль значения критических точек (показатель *t критическое двустороннее*) рассчитывается в ячейке D91 по формуле =СТЬЮДРАСПОБР(0,05;D86).

Напомним, что в подразд. 8.3 была рассмотрена функция ТТЕСТ, которая для двух выборочных массивов данных рассчитывает одностороннее или двустороннее *P*-значение *t*-теста. Если в данной функции аргумент *тип* = 1, то она выполняет *парный* двухвыборочный *t*-тест для средних.

Так, в рассмотренном примере функция ТТЕСТ может использоваться для расчета *одностороннего* (ячейка D88 в табл. 10.2) и *двустороннего* (ячейка D90 в табл. 10.2) *P*-значения парного

Таблица 10.2

	C	D	E
78	Парный двухвыборочный *t*-тест для средних		
79			
80		*Измерение при низкой температуре, см*	*Измерение при высокой температуре, см*
81	Среднее	10,40	10,41
82	Дисперсия	0,00044	0,00047
83	Наблюдения	10	10
84	Корреляция Пирсона	0,940	
85	Гипотетическая разность средних	0	
86	*df*	9	
87	*t*-статистика	– 3,86	
88	$P(T \le t)$ односторонняя	0,002	
89	*t* критическое одностороннее	1,83	
90	$P(T \le t)$ двусторонняя	0,004	
91	*t* критическое двустороннее	2,26	

t-теста. Это *P*-значение определяет уровень значимости, соответствующий расчетному критерию t_p и в режиме «Парный двухвыборочный *t*-тест для средних» вычисляется с помощью функции СТЬЮДРАСП.

Например, в табл. 10.2 двустороннее *P*-значение парного *t*-теста рассчитывается по формуле

=СТЬЮДРАСП(ABS(D87);D86;2),

которая адекватна формуле

= ТТЕСТ(D66 :D75;E66:E75;2;1).

РАЗДЕЛ III

Дисперсионный анализ

ГЛАВА 11
Однофакторный дисперсионный анализ

11.1. Краткие сведения из теории статистики

В главе 8 были рассмотрены процедуры оценки значимости различия между средними *двух* выборок. Первая из возможных вероятностных моделей строилась на предположении, что обе выборки извлечены из нормальных совокупностей с общей дисперсией ($\sigma_X^2 = \sigma_Y^2$), но, возможно, с различными математическими ожиданиями. С помощью этой модели проверялось, согласуются ли выборочные данные с нулевой гипотезой о фактическом равенстве этих математических ожиданий. На практике эти две выборки могли быть измерениями каких-либо сопоставимых величин, полученных в результате различных «обработок», а расхождение между математическими ожиданиями, если оно имеется, можно было приписать различию действия (эффекта) обработок. Например, измерения могли быть урожаями пшеницы, а две обработки соответствовали бы применению различных удобрений, так что одно из удобрений вносится на том поле, где собирают данные о первой выборке, а другое – на том, откуда поступают данные о второй выборке*.

Но как сравнить три обработки и более? Один из способов состоит в их попарном сравнении, когда для каждой пары применя-

*Следует заметить, что одними из первопроходцев в области разработки статистических методов проверки гипотез были исследователи, занимавшиеся изучением сельского хозяйства. Так, дисперсионный анализ первоначально был предложен Р. Фишером (1925) для обработки результатов агрономических опытов по выявлению условий, при которых испытываемый сорт сельскохозяйственной культуры дает максимальный урожай.

ются методы, рассмотренные в главе 8. Это довольно обременительно и не может быть признано удовлетворительным (не все пары будут независимыми), поэтому предпочтительнее обобщить двухвыборочную процедуру так, чтобы можно было ответить на вопрос: равны ли три (или более) математических ожидания?

Таким обобщением на три (и более) выборки является метод *дисперсионного анализа*, или ANOVA (Analysis of Variance – дисперсионный анализ), который служит для установления влияния отдельных факторов на изменчивость какого-либо признака, значения которого могут быть получены опытным путем в виде случайной величины Y. При этом величину Y называют *результативным признаком*, а конкретную реализацию фактора A – *уровнем* (*группой*) *фактора A* или *способом обра*ботки и обозначают через $A^{(i)}$.

В зависимости от числа оказывающих влияние факторов различают *однофакторный* и *многофакторный (двухфакторный* и т. д.) дисперсионный анализ.

Задачи однофакторного дисперсионного анализа хотя и являются самыми простыми в своем классе, но тем не менее весьма часто встречаются на практике. Типичный пример – сравнение по достигаемым результатам нескольких уровней фактора, например установление зависимости выполненных на стройке за смену работ от работающей бригады (*см.* пример 11.1).

Методы дисперсионного анализа основываются на идеях, во многом очень близких к тем, которые рассматривались в главах 8 и 9. Логика рассуждений при этом состоит в следующем.

Пусть $a_1, a_2,..., a_m$ – математическое ожидание результативного признака соответственно при уровне $A^{(1)}, A^{(2)},..., A^{(m)}$ ($i = 1, 2,..., m$).

Если при изменении уровня фактора групповые математические ожидания не изменяются, т. е. $a_1 = a_2 = ... = a_m$, то считаем, что результативный признак не зависит от фактора A, в противном случае такая зависимость имеется. Но поскольку числовые значения математических ожиданий неизвестны, возникает задача проверки гипотезы

$$H_0 : a_1 = a_2 = ... = a_m.$$

Проверить гипотезу о равенстве групповых математических ожиданий можно, соблюдая следующие требования при каждом уровне фактора:

1) наблюдения независимы и проводятся в одинаковых условиях;

2) результативный признак имеет нормальный закон распределения с постоянной для различных уровней генеральной дисперсией σ^2.

При этом возникает вопрос, как установить, одинаковы генеральные дисперсии результативного признака при различных уровнях фактора или нет? Не зная числовых значений этих дисперсий, нельзя однозначно ответить на этот вопрос, можно лишь проверить гипотезу

$$H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_m^2.$$

В главе 9 для проверки гипотезы $H_0 : \sigma_X^2 = \sigma_Y^2$ был предложен критерий Фишера, но он применим только для двух выборок. Для проверки гипотезы о равенстве дисперсий трех (и более) нормальных распределений применяется *критерий Бартлетта*

$$w = q\left[(n_1 - 1)\ln\frac{s^2}{s_1^2} + (n_2 - 1)\ln\frac{s^2}{s_2^2} + ... + (n_m - 1)\ln\frac{s^2}{s_m^2}\right],$$

где $q = \left[1 + \frac{1}{3(m-1)}\left(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} + ... + \frac{1}{n_m - 1} - \frac{1}{(n_1 - 1) + ... + (n_m - 1)}\right)\right]^{-1}$;

$$s^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2 + ... + (n_m - 1)s_m^2}{(n_1 - 1) + (n_2 - 1) + ... + (n_m - 1)};$$

$$s_i^2 = \frac{\tilde{\sigma}_i^2 n_i}{n_i - 1}, \quad i = 1, 2,..., m,$$

где $\tilde{\sigma}_i^2 = \frac{\sum_{j=1}^{n_i}(y_j^{(i)} - \bar{y}^{(i)})^2}{n_i}$ – групповая выборочная дисперсия;

n_i – численность наблюдений в группах.

При выполнении гипотезы $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_m^2$ величина w имеет распределение, близкое к χ^2-распределению с $k = m-1$ степенями свободы. Для подтверждения (или опровержения) гипотезы при заданном уровне значимости α находится правосторонняя критическая точка $w^{кр}_{пр,\alpha}$, определяющая критический интервал $(w^{кр}_{пр,\alpha}, +\infty)$. Если $w_р$ попадает в интервал $(w^{кр}_{пр,\alpha}, +\infty)$, то гипотеза $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_m^2$ отвергается, в противном случае – принимается.

Если гипотеза $H_0 : \sigma_1^2 = \sigma_2^2 = ... = \sigma_m^2$ подтверждается, то можно приступить непосредственно к процедуре дисперсионного анализа, т. е. к проверке гипотезы $H_0 : a_1 = a_2 = ... = a_m$. Сама процедура дисперсионного анализа базируется на том, что изменчивость или вариация наблюдаемых значений результативного признака Y может быть вызвана изменчивостью уровней фактора A и изменчивостью значений случайных неконтролируемых факторов, влияющих на Y, которые называют *остаточными*.

В математической статистике доказывается формула разложения общей выборочной дисперсии на сумму дисперсии групповых средних и средней из групповых дисперсий

$$\tilde{\sigma}_Y^2 = \tilde{\sigma}_{\bar{Y}(i)}^2 + \overline{\sigma_i^2} \quad (\text{или } \tilde{\sigma}_Y^2 = \tilde{\sigma}_Ф^2 + \tilde{\sigma}_О^2),$$

где $\tilde{\sigma}_Y^2$ – *общая выборочная дисперсия* – показатель вариации наблюдаемых «игреков», вызванной влиянием на Y фактора A и остаточных факторов;

$\tilde{\sigma}_{\bar{Y}(i)}^2 = \tilde{\sigma}_Ф^2$ – *дисперсия групповых средних* – показатель вариации наблюдаемых «игреков», вызванной влиянием на Y фактора A;

$\overline{\tilde{\sigma}_i^2} = \tilde{\sigma}_О^2$ – *средняя групповых дисперсий* – показатель вариации наблюдаемых «игреков», вызванной влиянием на Y остаточных факторов.

На основе данного разложения для генеральной дисперсии σ^2 находят три несмещенные оценки: $s_О^2$, $s_Ф^2$ и s_Y^2. Причем $s_О^2$ является несмещенной оценкой в любом случае, а $s_Ф^2$ и s_Y^2 – только при выполнении гипотезы $H_0 : a_1 = a_2 = ... = a_m$, т. е. только в том случае, когда фактор A не влияет на результативный признак Y.

Проверка гипотезы H_0 о равенстве групповых математических ожиданий основывается на сравнении оценок $s_Ф^2$ и $s_О^2$. В математической статистике доказывается, что если гипотеза H_0 : $a_1 = a_2 = ... = a_m$ верна, то величина

$$F = \frac{s_Ф^2}{s_О^2}$$

имеет F-распределение с числом степеней свободы $k=m-1$ и $l=n-m$, т. е.

$$\frac{s_Ф^2}{s_О^2} = F(k = m-1, l = n-m).$$

При использовании F-критерия строится правосторонняя критическая область $(F^{кр}_{пр.\alpha}, +\infty)$. Это объясняется тем, что в дисперсионном анализе, как правило, числитель больше знаменателя $(s_Ф^2 > s_О^2)$. Если это не так, то считают, что наблюдения не подтверждают влияние фактора на признак.

Если расчетное значение F-критерия $F_р$ попадает в интервал $(F^{кр}_{пр.\alpha}, +\infty)$, то гипотеза H_0 о равенстве групповых математических ожиданий отвергается, т. е. считаем, что фактор A влияет на результативный признак Y. Если же $F_р < F^{кр}_{пр.\alpha}$, то гипотеза H_0 не отвергается, и в этом случае говорят, что влияние фактора A на признак Y не подтвердилось выборочными наблюдениями.

Если в процессе анализа выявлено влияние фактора A на результативный признак Y, то можно измерить степень данного влияния с помощью *выборочного коэффициента детерминации*

$$\tilde{\rho}^2 = \frac{\tilde{\sigma}_Ф^2}{\tilde{\sigma}_Y^2},$$

который показывает, какая доля выборочной дисперсии $\tilde{\sigma}_Y^2$ объясняется зависимостью результативного признака Y от влияющего фактора A.

Итак, однофакторный дисперсионный анализ позволяет по выборочным данным выяснить, влияет ли контролируемый фактор на результативный признак, и при наличии такого влияния оценить его степень.

11.2. Справочная информация по технологии работы

Режим работы «Однофакторный дисперсионный анализ» служит для выяснения факта влияния контролируемого фактора *A* на результативный признак *Y* на основе выборочных данных.

В диалоговом окне данного режима (рис. 11.1) задаются следующие параметры (*см.* подразд. 1.1.2):

1. *Входной интервал.*
2. *Группирование.*
3. *Метки в первой строке/Метки в первом столбце.*
4. *Альфа* – вводится уровень значимости α, равный вероятности возникновения ошибки первого рода (отвержение нулевой гипотезы).
5. *Выходной интервал/Новый рабочий лист/Новая рабочая книга.*

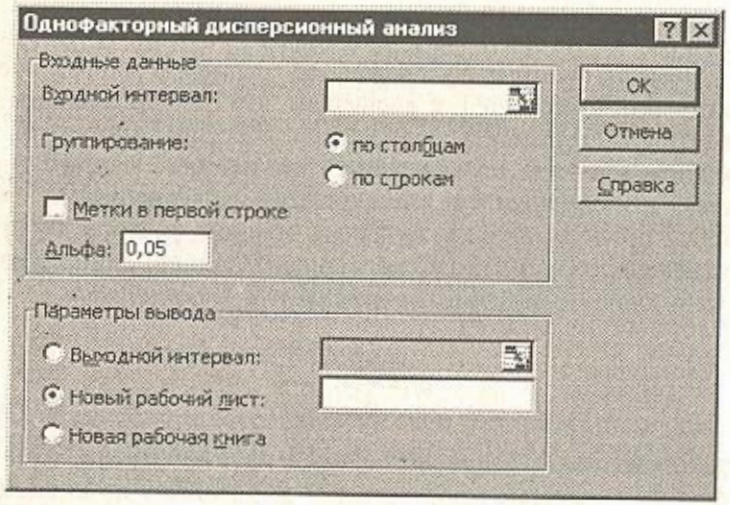

Рис. 11.1

Пример 11.1. Выборочные данные об объеме работ, выполненных на стройке (за смену) четырьмя бригадами, приведены в табл. 11.1, сформированной на рабочем листе Microsoft Excel [5].

Таблица 11.1

	B	C	D	E	F
3		Объем выполненной работы			
4	Номер смены	Бригада 1	Бригада 2	Бригада 3	Бригада 4
5	1	140	150	148	150
6	2	144	149	149	155
7	3	142	152	146	154
8	4	145	150	147	152

При уровне значимости $\alpha = 0{,}05$ требуется выяснить, зависит ли объем выполненных работ от работающей бригады.

Для решения задачи используем режим работы «Однофакторный дисперсионный анализ». Значения параметров, установленных в одноименном диалоговом окне, показаны на рис. 11.2.

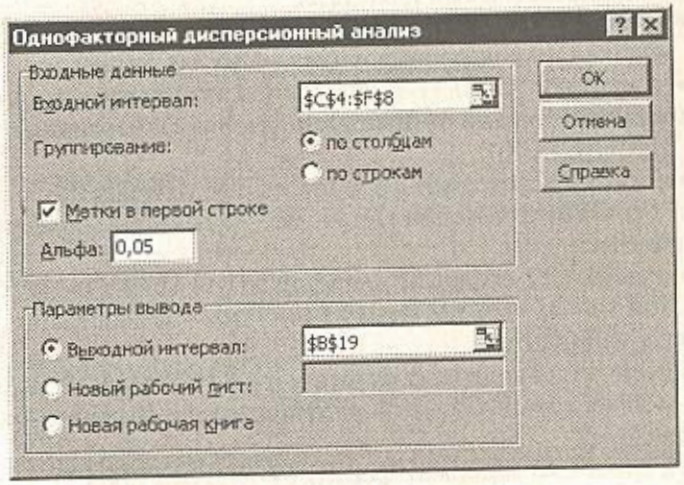

Рис. 11.2

Но прежде чем проводить анализ данных в сгенерированных таблицах, покажем, как с помощью критерия Бартлетта проверить гипотезу о равенстве генеральных дисперсий $H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \sigma_4^2$.

Показатели, рассчитанные в ходе проверки данной гипотезы, представлены в табл. 11.2.

Таблица 11.2

	B	C	D	E	F
9		Бригада 1	Бригада 2	Бригада 3	Бригада 4
10	Число наблюдений	4	4	4	4
11	Оценки s_i^2	4,92	1,58	1,67	4,92
12	Оценки s^2	3,27			
13	q	0,878			
14	w_p	1,540			
15	$w_{пр,\alpha}^{кр}$	7,81			

Содержимое ячеек в табл. 11.2:

- в массиве C10:F10 определяются объемы выборок n_i (например, ячейка C10 содержит формулу =СЧЕТ(C5:C8));
- в массиве C11:F11 вычисляются несмещенные оценки s_i^2 групповых дисперсий $\tilde{\sigma}_i^2$ (например, ячейка C11 содержит формулу =ДИСП(C5:C8));
- ячейка C12 содержит формулу {=СУММПРОИЗВ(C10:F10–1; C11:F11)/СУММ(C10:F10–1)} – рассчитывается объединенная оценка s^2;
- ячейка C13 содержит формулу {=1/(1+1/(3*(4–1))*(СУММ(1/(C10:F10–1))–1/СУММ(C10:F10–1)))} – вычисляется значение коэффициента q;
- ячейка C14 содержит формулу {=C13*СУММПРОИЗВ(C10:F10–1;LN(C12/C11:F11))} – рассчитывается значение критерия Бартлетта w_p;
- ячейка C16 содержит формулу =ХИ2ОБР(0,05;3) – определяется значение правосторонней критической точки $w_{пр,\alpha}^{кр}$.

Так как $w_p = 1{,}540$ не попадает в критическую область $(7{,}81; +\infty)$, то гипотеза $H_0 : \sigma_1^2 = \sigma_2^2 = \ldots = \sigma_m^2$ принимается и можно приступить к проверке гипотезы $H_0 : a_1 = a_2 = \ldots = a_m$. Показатели, рассчитанные в ходе проверки данной гипотезы, представлены в табл. 11.3 и 11.4.

Таблица 11.3

	B	C	D	E	F
19	**Однофакторный дисперсионный анализ**				
20					
21	ИТОГИ				
22	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
23	Бригада 1	4	571	142,75	4,92
24	Бригада 2	4	601	150,25	1,58
25	Бригада 3	4	590	147,5	1,67
26	Бригада 4	4	611	152,75	4,92

Таблица 11.4

	B	C	D	E	F	G	H
29	**Дисперсионный анализ**						
30	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-значение*	*F критическое*
31	Между группами	220,19	3	73,40	22,44	3,28E-0,5	3,49
32	Внутри групп	39,25	12	3,27			
33							
34	Итого	259,44	15				

Табл. 11.4 называется *таблицей однофакторного дисперсионного анализа*. Как видим, расчетное значение F-критерия $F_p = 22{,}44$, а

критическая область образуется правосторонним интервалом (3,49; +∞). Так как F_p попадает в критическую область, то гипотезу H_0 о равенстве групповых математических ожиданий отвергаем, т.е. считаем, что объем ежедневной выборки зависит от работающей бригады.

Выборочный коэффициент детерминации

$$\tilde{\rho}^2 = \frac{\tilde{\sigma}_\Phi^2}{\tilde{\sigma}_Y^2} = \frac{220{,}19/16}{259{,}44/16} \approx 0{,}85$$

показывает, что 85% общей выборочной вариации ежедневного объема выработки связано с работающей бригадой.

Рассмотрим более подробно механизм расчета основных показателей, представленных в табл. 11.4.

В ячейке C31 (показатель *SS между группами*) рассчитывается взвешенная сумма квадратов отклонений групповых средних от общей выборочной средней:

$$S_\Phi^2 = \sum_{i=1}^{m} (\overline{y}^{(i)} - \overline{y})^2 n_i.$$

В ячейке C32 (показатель *SS внутри групп*) вычисляется остаточная сумма квадратов отклонений наблюдаемых значений уровня от своей выборочной средней:

$$S_O^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_j^{(i)} - \overline{y}^{(i)})^2.$$

В ячейке C33 (показатель *SS итого*) рассчитывается общая сумма квадратов отклонений наблюдаемых значений от общей выборочной средней:

$$S_Y^2 = \sum_{i=1}^{m} \sum_{j=1}^{n_i} (y_j^{(i)} - \overline{y})^2 \text{ или } S_Y^2 = S_\Phi^2 + S_O^2.$$

В ячейках D31:D33 (показатель *df*) определяются *степени свободы*:

$$k_\Phi = m - 1 = 4 - 1 = 3;$$
$$k_O = n - m = 16 - 4 = 12;$$
$$k_Y = (m - 1) + (n - m) = n - 1 = 16 - 1 = 15.$$

В ячейках E31:E32 (показатель *MS*) вычисляются несмещенные оценки s_Φ^2 и s_O^2:

$$s_\Phi^2 = \frac{S_\Phi^2}{k_\Phi} = \frac{220{,}19}{3} \approx 73{,}40;$$

$$s_O^2 = \frac{S_O^2}{k_O} = \frac{39{,}25}{12} \approx 3{,}27.$$

В ячейке F31 (показатель *F*) вычисляется расчетное значение критерия F_p:

$$F_p = \frac{s_\Phi^2}{s_O^2} = \frac{73{,}40}{3{,}27} \approx 22{,}44.$$

В ячейке G31 (показатель *P-значение*) определяется *P-значение*, соответствующее расчетному значение критерия F_p, с помощью формулы

=FРАСП(F31;D31;D32).

В ячейке H31 (показатель *F критическое*) рассчитывается значение правосторонней критической точки $F^{кр}_{пр,\alpha}$ с помощью формулы

=FРАСПОБР(0,05;D31;D32).

ГЛАВА 12
Двухфакторный дисперсионный анализ без повторений и с повторениями

12.1. Краткие сведения из теории статистики

Продолжая тему главы 11, в которой была рассмотрена процедура однофакторного дисперсионного анализа, перейдем к задаче о действии на результативный признак Y двух факторов – A и B. Такие задачи характерны как для промышленных и технологических экспериментов, так и для гуманитарных исследований. Типичный пример – выяснение зависимости качества пряжи от типа станка и вида сырья, из которой она изготавливается (*см.* пример 12.1).

Логика однофакторного и двухфакторного дисперсионного анализа во многом схожа и состоит в следующем.

Пусть a_i – математическое ожидание результативного признака Y при уровне $A^{(i)}$ ($i = 1, 2, ..., m_A$); b_j – математическое ожидание результативного признака Y при уровне $B^{(j)}$ ($i = 1, 2, ..., m_B$). Если при изменении уровня фактора A групповые математические ожидания не изменяются, т. е. $a_1 = a_2 = ... = b_{m_A}$, то считаем, что результативный признак не зависит от фактора A, в противном случае такая зависимость имеется. Аналогично, если при изменении уровня фактора B сохраняется равенство $b_1 = b_2 = ... = b_{m_B}$, то считаем, что Y не зависит от фактора B. Но поскольку числовые значения математических ожиданий неизвестны, возникает задача проверки следующих гипотез:

$$H_A : a_1 = a_2 = ... = a_{m_A};$$

$$H_B : b_1 = b_2 = ... = b_{m_B}.$$

Проверять эти гипотезы, так же как и в задаче однофакторного дисперсионного анализа, можно только при соблюдении следующих требований:

1) при различных сочетаниях уровней факторов A и B наблюдения независимы;

2) при каждом сочетании уровней факторов A и B результативный признак Y имеет нормальный закон распределения с постоянной для различных сочетаний генеральной дисперсией σ^2.

Основой проведения двухфакторного дисперсионного анализа служит комбинационная группировка по двум факторам с последующим разложением дисперсии результативного признака $\tilde{\sigma}_Y^2$ по формуле

$$\tilde{\sigma}_Y^2 = \tilde{\sigma}_A^2 + \tilde{\sigma}_B^2 + \tilde{\sigma}_O^2,$$

где $\tilde{\sigma}_Y^2$ – *общая выборочная дисперсия* – показатель вариации наблюдаемых «игреков», вызванной влиянием на Y фактора A, фактора B и остаточных факторов;

$\tilde{\sigma}_A^2$ – *дисперсия групповых средних по фактору A* – показатель вариации наблюдаемых «игреков», вызванной влиянием на Y фактора A;

$\tilde{\sigma}_B^2$ – *дисперсия групповых средних по фактору B* – показатель вариации наблюдаемых «игреков», вызванной влиянием на Y фактора B;

$\tilde{\sigma}_O^2$ – *средняя групповых дисперсий* – показатель вариации наблюдаемых «игреков», вызванной влиянием на Y остаточных факторов.

На основе данного разложения для генеральной дисперсии σ^2 находятся четыре несмещенные оценки: s_O^2, s_A^2, s_B^2 и s_Y^2. Причем оценка s_O^2 является несмещенной оценкой в любом случае, оценка s_A^2 – при выполнении гипотезы $H_A : a_1 = a_2 = ... = a_{m_A}$, оценка s_B^2 – при выполнении гипотезы $H_B : b_1 = b_2 = ... = b_{m_B}$, а оценка s_Y^2 – при выполнении гипотез H_A и H_B.

Проверка гипотезы H_A основывается на сравнении дисперсий s_A^2 и s_O^2. В математической статистике доказывается, что если гипотеза H_A верна, то величина

$$F^A = \frac{s_A^2}{s_O^2}$$

имеет F-распределение с числом степеней свободы $k = m_A - 1$ и $l=(m_A-1)\,(m_B-1)$, т. е.

$$\frac{s_A^2}{s_O^2} = F(k = m_A - 1, l = (m_A - 1)(m_B - 1)).$$

Аналогичным образом рассчитывается и величина F^B.

Проверка выдвинутых гипотез осуществляется так же, как и при однофакторном дисперсионном анализе, и состоит в нахождении правосторонних критических интервалов $(F^{кр}_{пр,\alpha}, +\infty)$ с последующим контролем попадания (или непопадания) в данный интервал расчетных значений $F^A_р$ (или $F^B_р$). Если расчетное значение попадает в критический интервал, то гипотеза H_A (H_B) отвергается, т.е. считается, что фактор A (B) влияет на результативный признак Y.

Двухфакторный дисперсионный анализ может иметь две разновидности: без повторений и с повторениями. В первом случае каждому уровню факторов соответствует только одна выборка данных, во втором – определенным уровням факторов может соответствовать более одной выборки данных.

12.2. Справочная информация по технологии работы

Режимы работы «Двухфакторный дисперсионный анализ без повторений» и «Двухфакторный дисперсионный анализ с повторениями» служат для выяснения на основе выборочных данных факта влияния контролируемых факторов A и B на результативный признак Y. При этом в режиме «Двухфакторный дисперсионный анализ без повторений» каждому уровню факторов A и B соответствует *только одна* выборка данных, а в режиме «Двухфакторный дисперсионный анализ с повторениями» каждому уровню одного из факторов A (или B) соответствует *более одной* выборки данных. В последнем случае число выборок для каждого уровня должно быть *одинаковым*.

В диалоговых окнах данных режимов (рис. 12.1 и 12.2) задаются те же параметры, что и в диалоговом окне **Однофакторный дисперсионный анализ** (*см.* рис. 11.1), только добавлено поле *Число строк для выборки*. В это поле вводится число выборок, приходящихся на каждый уровень одного из факторов. Каждый

уровень фактора должен содержать одно и то же количество выборок (строк таблицы).

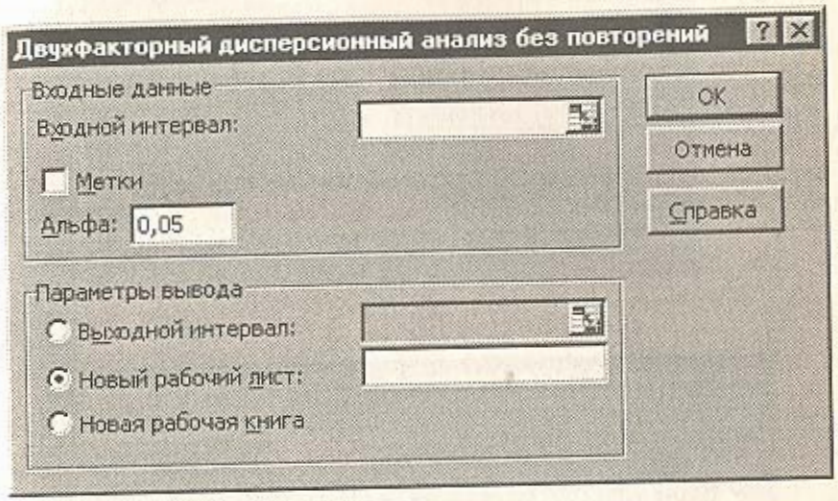

Рис. 12.1

Пример 12.1. Выборочные данные о разрывной нагрузке пряжи, изготовленной на разных станках и из отличающегося некоторым образом друг от друга сырья, приведены в табл. 12.1, сформированной на рабочем листе Microsoft Excel [5].

Таблица 12.1

	B	C	D
4	Тип станка	Вид сырья	
5		Шелк натуральный	Шелк искусственный
6	JANOME	10	50
7	HUSQVARNA	20	60
8	SINGER	30	100

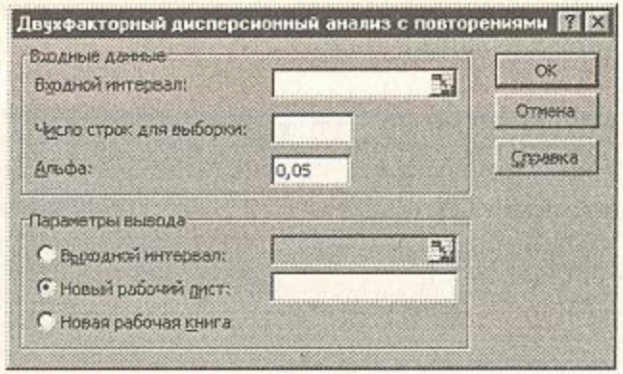

Рис. 12.2

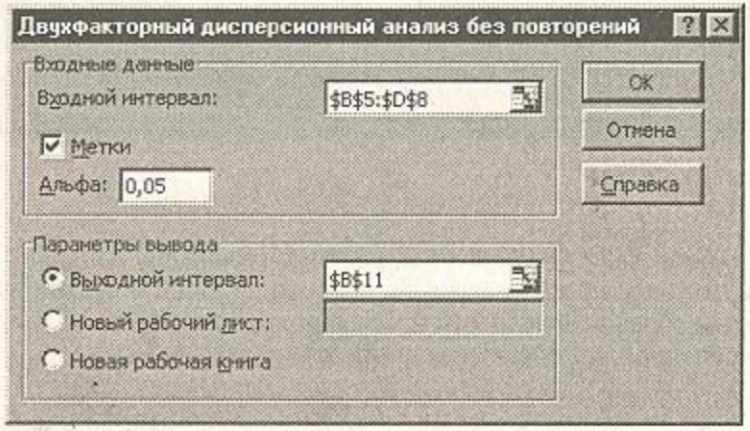

Рис. 12.3

Таблица 12.2

	B	C	D	E	F
11	Двухфакторный дисперсионный анализ без повторений				
12					
13	ИТОГИ	Счет	Сумма	Среднее	Дисперсия
14	JANOME	2	60	30	800
15	HUSQVARNA	2	80	40	800

Продолжение

	B	C	D	E	F
11	Двухфакторный дисперсионный анализ без повторений				
12					
13	ИТОГИ	Счет	Сумма	Среднее	Дисперсия
16	SINGER	2	130	65	2450
17					
18	Шелк натуральный	3	60	20	100
19	Шелк искусственный	3	210	70	700

Требуется при уровне значимости $\alpha = 0{,}05$ выяснить, влияют ли на качество пряжи, измеряемое величиной разрывной нагрузки, тип станка и вид сырья, из которого пряжа производится.

Для решения задачи используем режим работы «Двухфакторный дисперсионный анализ без повторений». Значения параметров, установленных в одноименном диалоговом окне, представлены на рис. 12.3, а рассчитанные в данном режиме показатели – в табл. 12.2 и 12.3.

Таблица 12.3

	B	C	D	E	F	G	H
22		Дисперсионный анализ					
23	Источник вариации	*SS*	*df*	*MS*	*F*	*P*-значение	*F* критическое
24	Строки	1300	2	650	4,33	0,187	19,00
25	Столбцы	3750	1	3750	25	0,038	18,51
26	Погрешность	300	2	150			
27							
28	Итого	5350	5				

Табл. 12.3 является *таблицей двухфакторного дисперсионного анализа без повторений*. Как видим, расчетное значение F-критерия фактора A (тип станка) $F_{\text{р}}^{A} = 4{,}33$, а критическая область образуется правосторонним интервалом (19,00; +∞). Так как $F_{\text{р}}^{A}$ не попадает в критическую область, то гипотезу $H_A : a_1 = a_2 = \ldots = a_{m_A}$ принимаем, т.е. считаем, что влияние типа станков на качество пряжи не подтвердилось.

Расчетное значение F-критерия фактора B (вид сырья) $F_{\text{р}}^{B} = 25$, а критическая область образуется правосторонним интервалом (18,51; +∞). Так как $F_{\text{р}}^{B}$ попадает в критическую область, то гипотезу $H_B : b_1 = b_2 = \ldots = b_{m_B}$ отвергаем, т.е. считаем, что вид сырья влияет на качество пряжи.

Выборочный коэффициент детерминации

$$\tilde{\rho}_B^2 = \frac{\tilde{\sigma}_B^2}{\tilde{\sigma}_Y^2} = \frac{3750/6}{5350/6} \approx 0{,}70$$

показывает, что 70 % общей выборочной вариации качества пряжи связано с влиянием на нее вида сырья.

Механизмы расчета показателей, представленных в табл. 12.3 и 11.4 (*см.* подразд. 11.2), во многом схожи.

Рассмотрим технологию работы в режиме «Двухфакторный дисперсионный анализ с повторениями».

Пример 12.2. Выборочные данные об урожайности пшеницы, выращенной на участках, на которые вносились различные виды удобрений и которые подвергались различной химической обработке, приведены в табл. 12.4, сформированной на рабочем листе Microsoft Excel [6].

Требуется при уровне значимости $\alpha = 0{,}05$ выяснить, влияют ли на урожайность пшеницы вид удобрения и способ химической обработки почвы.

Рассматриваемый в задаче эксперимент представляет собой факторный эксперимент типа 4×4, при котором четыре вида удобрений (фактор A) пересекаются с использованием четырех способов химической обработки почвы (фактор B). Таким образом, в плане эксперимента имеется 16 условий. Но в отличие от ранее рассмотренной задачи (*см.* пример 12.1) здесь каждому условию соответствует не одно, а три значения (3 участка земли, засеянных пшеницей).

Таблица 12.4

	B	C	D	E	F	G
32	Номер участка	Вид удобрения	Способ химической обработки			
33			Способ 1	Способ 2	Способ 3	Способ 4
34	Участок 1	Удобрение 1	21,4	20,9	19,6	17,6
35	Участок 2		21,2	20,3	18,8	16,6
36	Участок 3		20,1	19,8	16,4	17,5
37	Участок 1	Удобрение 2	12,0	13,6	13,0	13,3
38	Участок 2		14,2	13,3	13,7	14,0
39	Участок 3		12,1	11,6	12,0	13,9
40	Участок 1	Удобрение 3	13,5	14,0	12,9	12,4
41	Участок 2		11,9	15,6	12,9	13,7
42	Участок 3		13,4	13,8	12,1	13,0
43	Участок 1	Удобрение 4	12,8	14,1	14,2	12,0
44	Участок 2		13,8	13,2	13,6	14,6
45	Участок 3		13,7	15,3	13,3	14,0

Для решения задачи используем режим работы «Двухфакторный дисперсионный анализ с повторениями». Значения параметров, установленных в одноименном диалоговом окне, представлены на рис. 12.4, а рассчитанные в данном режиме показатели — в табл. 12.5 и 12.6.

Табл. 12.6 является *таблицей двухфакторного дисперсионного анализа с повторениями*. Как видим, расчетное значение F-критерия фактора A (вид удобрения) $F_{\text{р}}^{A} = 123{,}64$, а критическая область образуется правосторонним интервалом (2,90; +∞).

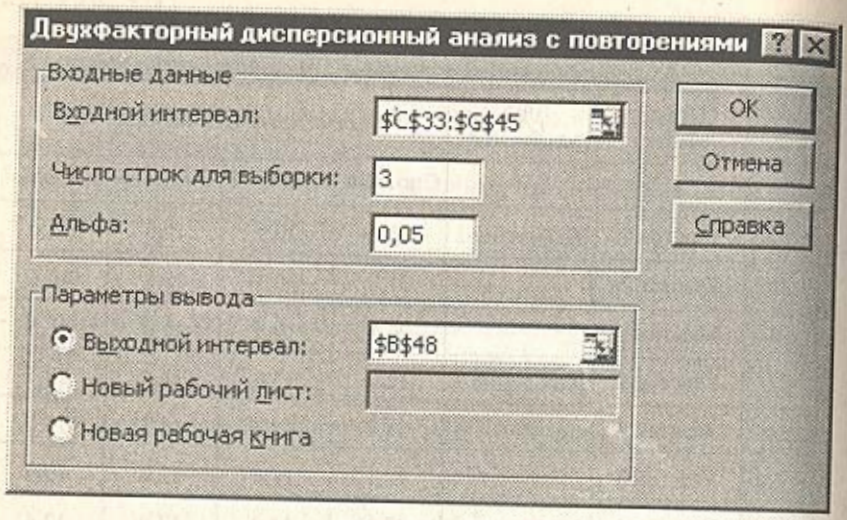

Рис. 12.4

Таблица 12.5

	B	C	D	E	F	G
48	**Двухфакторный дисперсионный анализ с повторениями**					
49						
50	ИТОГИ	Способ 1	Способ 2	Способ 3	Способ 4	Итого
51	*Удобрение 1*					
52	Счет	3	3	3	3	12
53	Сумма	62,7	61	54,8	51,7	230,2
54	Среднее	20,90	20,33	18,27	17,23	19,18
55	Дисперсия	0,49	0,30	2,77	0,30	3,13
56						
57	*Удобрение 2*					
58	Счет	3	3	3	3	12
59	Сумма	38,3	38,5	38,7	41,2	156,7

Продолжение

	B	C	D	E	F	G
60	Среднее	12,77	12,83	12,90	13,73	13,06
61	Дисперсия	1,54	1,16	0,73	0,14	0,82
62						
63	*Удобрение 3*					
64	Счет	3	3	3	3	12
65	Сумма	38,8	43,4	37,9	39,1	159,2
66	Среднее	12,93	14,47	12,63	13,03	13,27
67	Дисперсия	0,80	0,97	0,21	0,42	0,99
68						
69	*Удобрение 4*					
70	Счет	3	3	3	3	12
71	Сумма	40,3	42,6	41,1	40,6	164,6
72	Среднее	13,43	14,20	13,70	13,53	13,72
73	Дисперсия	0,30	1,11	0,21	1,85	0,73
74						
75	*Итого*					
76	Счет	12	12	12	12	
77	Сумма	180,1	185,5	172,5	172,6	
78	Среднее	15,01	15,46	14,38	14,38	
79	Дисперсия	13,26	9,71	6,39	3,52	

Таблица 12.6

	B	C	D	E	F	G	H
82	Дисперсионный анализ						
83	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-значение*	*F критическое*
84	Выборка	309,26	3	103,09	123,64	1,11E-17	2,90
85	Столбцы	9,97	3	3,32	3,99	0,016	2,90
86	Взаимодействие	25,68	9	2,85	3,42	0,005	2,19
87	Внутри	26,68	32	0,83			
88							
89	Итого	371,59	47				

Так как $F_{\text{р}}^{A}$ попадает в критическую область, то гипотезу H_A отвергаем, т. е. считаем, что вид удобрения влияет на урожайность пшеницы.

Выборочный коэффициент детерминации для фактора A

$$\tilde{\rho}_A^2 = \frac{\tilde{\sigma}_A^2}{\tilde{\sigma}_Y^2} = \frac{309{,}26/48}{371{,}59/48} \approx 0{,}83$$

показывает, что 83 % общей выборочной вариации урожайности пшеницы связано с влиянием вида удобрения.

Расчетное значения F-критерия фактора B (способ химической обработки) $F_{\text{р}}^{B} = 3{,}99$, а критическая область образуется правосторонним интервалом $(2{,}90; +\infty)$. Так как $F_{\text{р}}^{B}$ попадает в критическую область, то гипотезу H_B отвергаем, т. е. считаем, что способ химической обработки почвы также влияет на урожайность пшеницы.

Выборочный коэффициент детерминации для фактора B

$$\tilde{\rho}_B^2 = \frac{\tilde{\sigma}_B^2}{\tilde{\sigma}_Y^2} = \frac{9{,}97/48}{371{,}59/48} \approx 0{,}03$$

показывает, что только около 3 % общей выборочной вариации урожайности пшеницы связано с влиянием способа химической обработки почвы.

Значимость фактора взаимодействия $F_{\text{р}}^{AB}$ ($F_{\text{р}}^{AB} = 3{,}42$ и попадает в критический интервал $(2{,}19; +\infty)$) указывает на то, что эффективность различных видов удобрения варьируется при различных способах химической обработки почвы.

Механизмы расчета показателей, представленных в табл. 12.6 и 11.4 (*см.* подразд. 11.2), во многом аналогичны.

РАЗДЕЛ IV

Статистические методы изучения взаимосвязей явлений и процессов

ГЛАВА 13
Ковариация и корреляция

13.1. Краткие сведения из теории статистики

В экономических исследованиях одной из важных задач является анализ зависимостей между изучаемыми переменными. Зависимость между переменными может быть либо *функциональной*, либо *стохастической* (*вероятностной*). Для оценки тесноты и направления связи между изучаемыми переменными при их стохастической зависимости пользуются показателями *ковариации* и *корреляции*.

Ковариацией $cov(x, y)$ случайных величин X и Y называют среднее произведений отклонений каждой пары значений величин X и Y в исследуемых массивах данных:

$$\operatorname{cov}(x,y) = \overline{(x_i - \bar{x})(y_i - \bar{y})} = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y}).$$

Ковариация есть характеристика системы случайных величин, описывающая помимо рассеивания величин X и Y еще и линейную связь между ними. Доказано [1], что для независимых случайных величин X и Y их ковариация равна нулю, а для зависимых случайных величин она отличается от нуля (хотя и не обязательно). Поэтому ненулевое значение ковариации означает зависимость случайных величин. Однако обращение в нуль ковариации не гарантирует независимости, бывают зависимые случайные величины, ковариация которых равна нулю. Из формулы определения ковариации видно, что ковариация характеризует не только зависимость величин, но и их рассеивание. Действительно, если, например, одна из величин X или Y мало отличается от своего математического ожидания (почти не случайна), то показатель ковариации будет мал, какой бы тесной зависимостью ни были связаны величины X и Y. Так что обращение в нуль ковариации величин X и Y является не достаточным условием для их независимости, а только необходимым.

Использование ковариации в качестве меры связи признаков не совсем удобно, так как показатель ковариации не нормирован и при переходе к другим единицам измерения (например, от метров к километрам) меняет значение. Поэтому в статистическом анализе показатель ковариации сам по себе используется редко; он фигурирует обычно как промежуточный элемент расчета линейного коэффициента корреляции r_{xy}:

$$r_{xy} = \frac{\operatorname{cov}(x,y)}{\sigma_x \sigma_y}.$$

В 1889 г. Ф. Голтон* высказал мысль о коэффициенте, который мог бы измерить тесноту связи между двумя коррелируемыми признаками. В начале 90-х гг. XIX в. Пирсон, Эджворт и Велдон получили формулу линейного коэффициента корреляции

$$r_{xy} = \frac{\overline{xy} - \bar{x}\,\bar{y}}{\sigma_x \sigma_y}.$$

Линейный коэффициент корреляции характеризует степень тесноты не всякой, а только линейной зависимости. При нелиней-

*(Galton Francis) Голтон Фрэнсис (1822–1911) – английский психолог и антрополог. В математике Голтон разработал методы статистической обработки результатов исследований (в частности, метод исчисления корреляций между переменными); ввел коэффициент корреляции; создал так называемую биометрическую школу.

ной зависимости между явлениями линейный коэффициент корреляции теряет смысл, и для измерения тесноты связи применяют так называемое корреляционное отношение, известное также под названием «индекс корреляции» [9, 12].

Линейная вероятностная зависимость случайных величин заключается в том, что при возрастании одной случайной величины другая имеет тенденцию возрастать (или убывать) по линейному закону. Эта тенденция к линейной зависимости может быть более или менее ярко выраженной, т. е. более или менее приближаться к функциональной. Если случайные величины X и Y связаны точной линейной функциональной зависимостью $y=ax+b$, то $r_{xy} = \pm 1$. В общем случае, когда величины X и Y связаны произвольной вероятностной зависимостью, линейный коэффициент корреляции принимает значение в пределах $-1 < r_{xy} < 1$, тогда качественная оценка тесноты связи величин X и Y может быть выявлена на основе шкалы Чеддока (табл. 13.1).

Таблица 13.1

Теснота связи	Значение коэффициента корреляции при наличии:	
	прямой связи	обратной связи
Слабая	0,1 – 0,3	(–0,1) – (–0,3)
Умеренная	0,3 – 0,5	(–0,3) – (–0,5)
Заметная	0,5 – 0,7	(–0,5) – (–0,7)
Высокая	0,7 – 0,9	(–0,7) – (–0,9)
Весьма высокая	0,9 – 0,99	(–0,9) – (–0,99)

В теории разработаны и на практике применяются различные модификации формул расчета линейного коэффициента корреляции:

$$r_{xy} = \frac{\sum(x-\bar{x})(y-\bar{y})}{n\sigma_x\sigma_y}; \qquad (13.1)$$

$$r_{xy} = \frac{n\sum xy - \sum x\sum y}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}; \qquad (13.2)$$

$$r_{xy} = \frac{\sigma_x^2 + \sigma_y^2 - \sigma_{x-y}^2}{2\sigma_x\sigma_y}. \qquad (13.3)$$

Приведенные формулы в определенных случаях имеют некоторые преимущества друг перед другом. Например, при небольших значениях n ($n < 30$) обычно употребляются формулы (13.2) и (13.3).

Необходимо обратить внимание, что формулы (13.1) – (13.3) справедливы для нахождения *генерального* коэффициента корреляции. Чтобы рассчитать *выборочный* коэффициент корреляции, необходимо в этих формулах генеральные средние заменить на выборочные средние, а генеральные стандартные отклонения – на выборочные стандартные отклонения.

13.2. Справочная информация по технологии работы

Режим работы «Ковариация» служит для расчета *генеральной* ковариации на основе *выборочных* данных.

Режим работы «Корреляция» предназначен для расчета *генерального* и *выборочного* коэффициентов корреляции соответственно на основе генеральных и выборочных данных.

В диалоговых окнах данных режимов (рис. 13.1 и 13.2) задаются параметры, аналогичные параметрам, задаваемым в диалоговом окне **Ранг и персентиль** (*см.* рис. 5.1).

Пример 13.1. Показатели уровня образования, уровня преступности, а также отношение числа безработных к числу вакансий в некоторых центральных областях России в 1995 г. (по данным Госкомстата РФ) приведены в табл. 13.2, сформированной на рабочем листе Microsoft Excel.

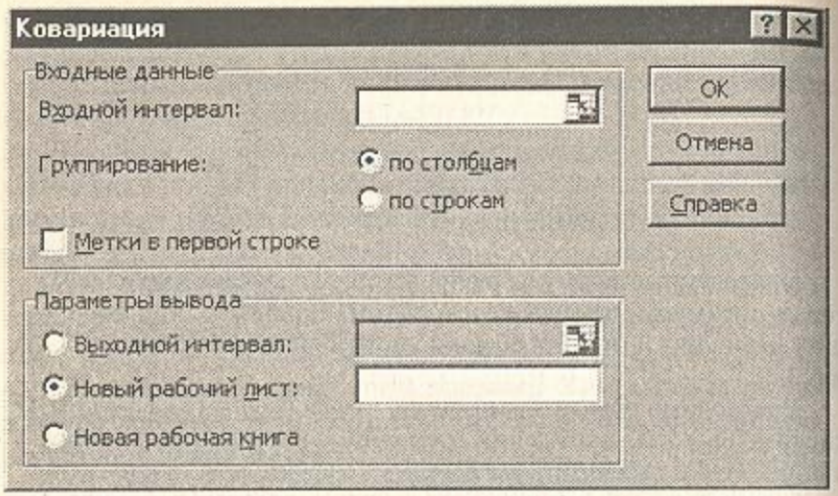

Рис. 13.1

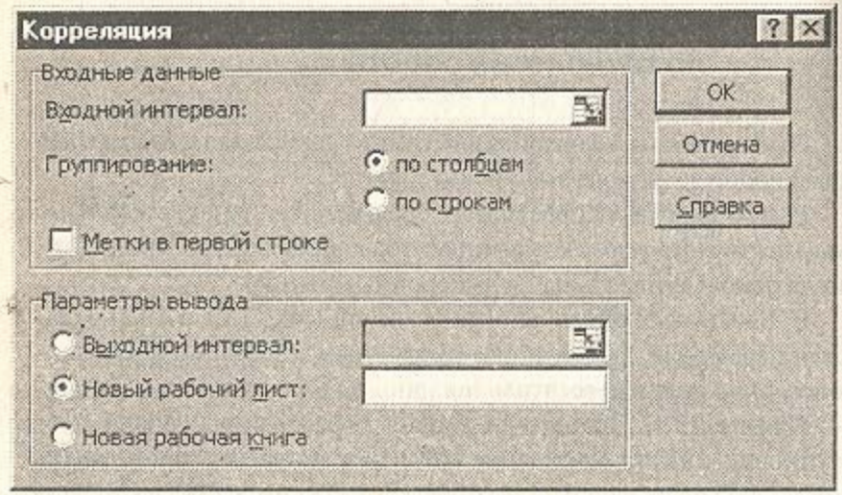

Рис. 13.2

По выборочным данным, представленным в табл. 13.2, требуется установить наличие взаимосвязи между указанными показателями в центральном регионе России.

Таблица 13.2

	B	C	D	E
4	Область	Уровень образования	Отношение числа безработных к числу вакансий	Уровень преступности
5	Брянская	735	22,3	908
6	Владимирская	788	10,8	791
7	Ивановская	779	52,9	804
8	Калужская	795	2,2	701
9	Костромская	740	10,4	685
10	г. Москва	902	0,4	496
11	Московская	838	2,4	536
12	Нижегородская	763	5,4	936
13	Орловская	762	4,1	662
14	Рязанская	757	4,1	671
15	Смоленская	772	1,0	920
16	Тверская	764	4,2	1040
17	Тульская	764	2,1	809
18	Ярославская	755	25,1	882

Примечания: 1. Уровень образования рассчитывался как численность лиц с высшим и средним специальным образованием на 1000 жителей области.

2. Уровень преступности рассчитывался как число совершенных преступлений на 100 тыс. жителей области.

Для решения задачи используем режимы работы «Ковариация» и «Корреляция». Значения параметров, установленных в одноименных диалоговых окнах, представлены на рис. 13.3 и 13.4, а рассчитанные в данных режимах показатели – в табл. 13.3 и 13.4.

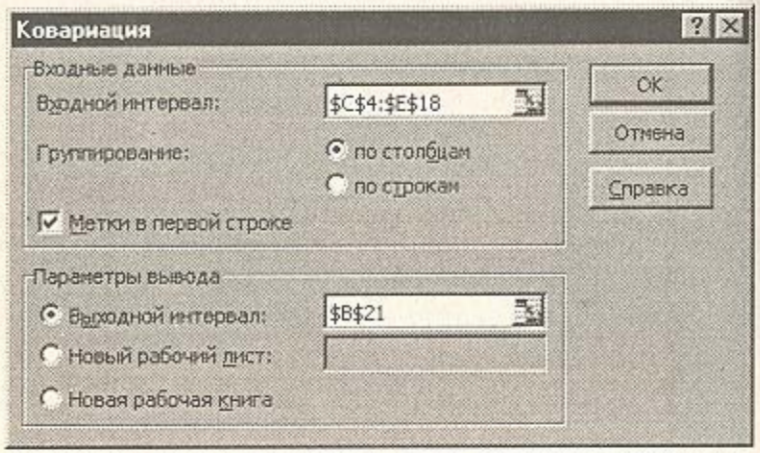

Рис. 13.3

Таблица 13.3

	B	C	D	E
21		Уровень образования	Отношение числа безработных к числу вакансий	Уровень преступности
22	Уровень образования	1884,88		
23	Отношение числа безработных к числу вакансий	– 161,39	207,32	
24	Уровень преступности	– 4479,22	536,80	24667,63

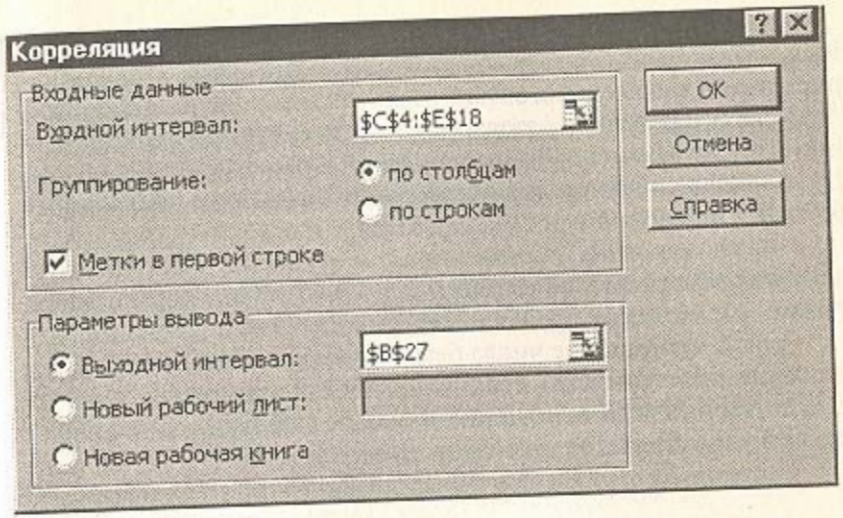

Рис. 13.4

Таблица 13.4

	B	C	D	E
27		Уровень образования	Отношение числа безработных к числу вакансий	Уровень преступности
28	Уровень образования	1		
29	Отношение числа безработных к числу вакансий	– 0,26	1	
30	Уровень преступности	0,66	0,24	1

Как видно из табл. 13.3 и 13.4, между парами всех исследуемых показателей существуют стохастические связи. Причем характер всех выявленных связей различен и состоит в следующем:

• связь «уровень образования» – «отношение числа безработных к числу вакансий» является слабой и обратной ($r_{xy} = -0{,}26$), т. е. с повышением уровня образования отношение числа безработных к числу вакансий уменьшается;

• связь «уровень образования» – «уровень преступности» является заметной и обратной ($r_{xy} = -0{,}66$), т. е. с повышением уровня образования уровень преступности уменьшается;

• связь «отношение числа безработных к числу вакансий» – «уровень преступности» является слабой и прямой ($r_{xy} = 0{,}24$), т. е. с увеличением отношения числа безработных к числу вакансий увеличивается и уровень преступности.

13.3. Статистические функции, связанные с режимами «Ковариация» и «Корреляция»

Функция КОВАР

См. также КОРРЕЛ, ФИШЕР, ФИШЕРОБР.

Синтаксис:

КОВАР (массив1; массив2)

Результат:

Рассчитывает значение ковариации между двумя массивами данных.

Аргументы:

• *массив1*: первый массив данных;

• *массив2*: второй массив данных.

Замечания:

• аргументы должны быть числами или массивами, содержащими числа;

• если аргумент, который является массивом, содержит текстовые, логические значения или пустые ячейки, то такие значения игнорируются, однако ячейки с нулевыми значениями учитываются;

• если аргументы *массив1* и *массив2* имеют различное количество точек данных, то функция КОВАР помещает в ячейку значение ошибки #Н/Д;

• если аргумент *массив1* либо *массив2* пуст, то функция КОВАР помещает в ячейку значение ошибки #ДЕЛ/0!.

Математико-статистическая интерпретация:

См. подразд. 13.1.

Примечание. В отличие от режима «Ковариация» функция КОВАР рассчитывает значение ковариации в предположении, что массивы данных образуют генеральные совокупности.

♦ В примере 13.1 (*см.* табл. 13.3) функция КОВАР совместно с функцией СЧЕТ используется для расчета показателей ковариации. Например, значение в ячейке С22 рассчитывается по формуле

=КОВАР(C5:C18;C5:C18)*СЧЕТ(C5:C18)/(СЧЕТ(C5:C18)–1),

а значение в ячейке С23 – по формуле

=КОВАР(C5:C18;D5:D18)*СЧЕТ(C5:C18)/(СЧЕТ(C5:C18)–1).

Функция КОРРЕЛ

См. также ПИРСОН, КОВАР, ФИШЕР, ФИШЕРОБР.

Синтаксис:

КОРРЕЛ (массив1; массив2)

Результат:

Рассчитывает линейный коэффициент корреляции между массивами данных.

Аргументы:

• *массив1*: первый массив данных;

• *массив2*: второй массив данных.

Замечания:

• аргументы должны быть числами или именами, массивами или ссылками, содержащими числа;

• если аргумент, который является массивом, содержит текстовые, логические значения или пустые ячейки, то такие значения игнорируются; однако ячейки с нулевыми значениями учитываются;

• если аргументы *массив1* и *массив2* имеют различное количество точек данных, то функция КОРРЕЛ помещает в ячейку значение ошибки #Н/Д;

• если аргумент *массив1* либо *массив2* пуст или если стандартное отклонение их значений равно 0, то функция КОРРЕЛ помещает в ячейку значение ошибки #ДЕЛ/0!.

Математико-статистическая интерпретация:

См. подразд. 13.1.

◆ В примере 13.1 (*см.* табл. 13.4) функция КОРРЕЛ используется для расчета коэффициентов корреляции между исследуемыми признаками. Например, значение в ячейке С28 рассчитывается по формуле

=КОРРЕЛ(C5:C18;C5:C18),

а значение в ячейке С29 – по формуле

=КОРРЕЛ(C5:C18;D5:D18).

13.4. Родственные статистические функции

В подразд. 13.3 были рассмотрены статистические функции КОВАР и КОРРЕЛ, используемые для расчетов соответственно в режимах «Ковариация» и «Корреляция». Здесь приводятся описания функций ФИШЕР и ФИШЕРОБР, родственных по своей сущности данным режимам.

Функция ФИШЕР

См. также ФИШЕРОБР, КОРРЕЛ, ПИРСОН, КОВАР.

Синтаксис:

ФИШЕР (*x*)

Результат:

Рассчитывает преобразование Фишера для аргумента *x*.

Аргументы:

x: числовое значение, которое необходимо преобразовать.

Замечания:

• если аргумент *x* не является числом, то функция ФИШЕР помещает в ячейку значение ошибки # ЗНАЧ!;

• если аргумент $x \le -1$ или аргумент $x \ge 1$, то функция ФИШЕР помещает в ячейку значение ошибки # ЧИСЛО!.

Математико-статистическая интерпретация:

На практике коэффициент корреляции, а также параметры уравнения регрессии (*см.* главу 14) определяются чаще всего по выборочным данным, следовательно, полученные выборочные показатели отличаются от аналогичных показателей в генеральной совокупности. В связи с этим необходимо определять точность показателей корреляции и границы доверительных интервалов.

Выборочный коэффициент корреляции r_{xy} представляет собой случайную величину, поэтому его распределение можно считать нормальным или приближенно нормальным, если выполняются следующие условия:

• переменные X и Y, между которыми определяется корреляционная связь, имеют совместное нормальное или приближенно нормальное распределение;

• коэффициент корреляции не равен ± 1;

• объем выборки достаточно велик.

При невыполнении указанных выше условий распределение коэффициента корреляции отличается от нормального. В этом случае для проверки гипотезы о наличии корреляционной связи, а также для построения доверительного интервала коэффициент корреляции преобразуют в величину z, имеющую приблизительно нормальное распределение и рассчитывающуюся по формуле

$$z = \frac{1}{2}\ln\frac{1+r}{1-r}.$$

Данное выражение получило название *«z-преобразования Фишера»**.

---

*(Fisher Ronald Aylmer) Фишер Роналд Эйлмер (1890–1962) – английский статистик и генетик, член Лондонского королевского общества (1929). Основные труды по статистике и генетической теории эволюции, построил теорию точечных и интервальных статистических оценок, разработал методику планирования экспериментов и внес существенный вклад в создание теории статистической проверки гипотез.

Пример 13.2. Требуется на основе выборочных данных о деловой активности однотипных коммерческих структур оценить тесноту связи между прибылью Y (млн руб.) и затратами X (руб.) на производство единицы продукции (диапазон B2:D8 в табл. 13.5) [12].

Таблица 13.5

	B	C	D
2	№ п/п	Y	X
3	1	221	96
4	2	1070	77
5	3	1001	77
6	4	606	89
7	5	779	82
8	6	789	81
9	Коэффициент корреляции r_{xy}		–0,984
10	Расчетное значение t-критерия $t_р$		11,011
11	Табличное значение t-критерия $t_{кр}$		2,776
12	Табличное значение стандартного нормального распределения z_γ		1,960
13	Значение преобразования Фишера z'		–2,407
14	Левая интервальная оценка для z		–3,539
15	Правая интервальная оценка для z		–1,275
16	Левая интервальная оценка для r_{xy}		–0,998
17	Правая интервальная оценка для r_{xy}		–0,855
18	Стандартное отклонение для r_{xy}		0,014

Общая схема решения подобных задач состоит в следующем:

1. По одной из формул (13.1) – (13.3) рассчитывается линейный коэффициент корреляции r_{xy}.

2. Проверяется значимость линейного коэффициента корреляции на основе t-критерия Стьюдента (*см.* описание функций СТЬЮДРАСП и СТЬЮДРАСПОБР в подразд. 6.3.8). При этом

выдвигается и проверяется гипотеза H_0 о равенстве коэффициента корреляции нулю ($H_0 : r_{xy} = 0$). При проверке этой гипотезы используется t-статистика:

$$t_р = \sqrt{\frac{r^2}{1-r^2}(n-2)} = \frac{|r|}{\sqrt{1-r^2}}\sqrt{n-2}.$$

Если гипотеза H_0 подтверждается, t-статистика имеет распределение Стьюдента с входными параметрами α и k (α – уровень значимости; $k = n - 2$ – число степеней свободы). Если расчетное значение $t_р > t_{кр}$, то гипотеза $H_0 : r_{xy} = 0$ отвергается, что свидетельствует о значимости линейного коэффициента корреляции, а следовательно, и о статистической существенности зависимости между X и Y.

3. Для статистически значимого линейного коэффициента корреляции определяется интервальная оценка для z по выражению

$$z \in \left[z' \pm z_\gamma \sqrt{\frac{1}{n-3}} \right],$$

где z' – значение, полученное на основе z-преобразования Фишера;

z_γ – табулированные значения для стандартного нормального распределения, зависимые от $\gamma = 1 - \alpha$;

n – размер выборочной совокупности.

4. На основе обратного z-преобразования Фишера определяется интервальная оценка для линейного коэффициента корреляции r_{xy}.

5. Рассчитывается стандартная ошибка линейного коэффициента корреляции по формуле

$$\sigma_\gamma = \sqrt{\frac{1-r^2}{n-2}}.$$

Результаты решения задачи приведены в табл. 13.5.

Содержимое ячеек в табл. 13.5:

- массив B3:D8 содержит исходные данные задачи;
- ячейка D9 содержит формулу =КОРРЕЛ(C3:C8;D3:D8) – рассчитывается значение линейного коэффициента корреляции (п. 1 общей схемы решения задачи);
- ячейка D10 содержит формулу =ABS(D9)/КОРЕНЬ(1-СТЕПЕНЬ(D9;2))*КОРЕНЬ(6–2) – вычисляется расчетное значение t-критерия $t_р$ (п. 2 общей схемы решения задачи);
- ячейка D11 содержит формулу =СТЬЮДРАСПОБР(0,05;4) – рассчитывается табличное значение t-критерия $t_{кр}$ ($\alpha = 0{,}05$; $k = n - 2 = 6 - 2 = 4$). Выполнение неравенства $t_р > t_{кр}$ свидетельствует о значимости линейного коэффициента корреляции;
- ячейка D12 содержит формулу =НОРМСТОБР((0,95+1)/2) – вычисляется табулированное значение стандартного нормального распределения (*см.* описание функций НОРМОБР и НОРМСТОБР в подразд. 6.3.1);
- ячейка D13 содержит формулу =ФИШЕР(D9) – определяется значение z', полученное на основе преобразования Фишера;
- ячейки D14 и D15 содержат формулы =D13–D12*КОРЕНЬ(1/(6 – 3)) и =D13+D12*КОРЕНЬ(1/(6–3)) – рассчитываются интервальные оценки z (п. 3 общей схемы решения задачи);
- ячейки D16 и D17 содержат формулы =ФИШЕРОБР(D14) и =ФИШЕРОБР(D15) – вычисляются интервальные оценки линейного коэффициента корреляции (п. 4 общей схемы решения задачи);
- ячейка D18 содержит формулу =КОРЕНЬ((1-D9^2)/(6 – 2)) – рассчитывается значение стандартной ошибки линейного коэффициента корреляции (п. 5 общей схемы решения задачи).

Таким образом, с вероятностью 0,95 линейный коэффициент корреляции заключен в интервале от –0,855 до –0,998 со стандартной ошибкой 0,014. Следовательно, прибыль обследованных коммерческих структур находится в тесной связи с затратами на производство единицы продукции.

Функция ФИШЕРОБР

См. также ФИШЕР, КОРРЕЛ, ПИРСОН, КОВАР.

Синтаксис:

ФИШЕРОБР (z)

Результат:

Рассчитывает обратное преобразование Фишера.

Аргументы:

z: значение, для которого осуществляется обратное преобразование Фишера.

Замечания:

если *z* не является числом, то функция ФИШЕРОБР помещает в ячейку значение ошибки #ЗНАЧ!.

Математико-статистическая интерпретация:

См. описание функции ФИШЕР.

Функция обратного преобразования Фишера используется в ситуациях, когда известно значение, полученное на основе прямого преобразования Фишера, и необходимо найти значение аргумента этого преобразования.

Например, формула =ФИШЕРОБР(–2,407) вычисляет значение –0,984 (сравните с формулой =ФИШЕР(–0,984), рассчитывающей значение –2,407 в ячейке D13 табл. 13.5).

Уравнение для обратного преобразования Фишера имеет следующий вид:

$$x = \frac{e^{2z} - 1}{e^{2z} + 1}.$$

ГЛАВА 14
Регрессия

14.1. Краткие сведения из теории статистики

В главе 13 были рассмотрены основные аспекты корреляционного анализа, который имеет своей задачей определение тесноты и направления связи между изучаемыми величинами. Наряду с корреляционным анализом обычно проводится и *регрессионный анализ*, который заключается в определении аналитического выражения связи зависимой случайной величины Y (называемой также *результативным признаком*) с независимыми случайными величинами $X_1, X_2, ..., X_m$ (называемыми также *факторами*).

Форма связи результативного признака Y с факторами X_1, X_2,..., X_m получила название *уравнения регрессии*. В зависимости от типа выбранного уравнения различают *линейную* и *нелинейную* регрессию (в последнем случае возможно дальнейшее уточнение: квадратичная, экспоненциальная, логарифмическая и т. д.).

В зависимости от числа взаимосвязанных признаков различают *парную* и *множественную* регрессию. Если исследуется связь между двумя признаками (результативным и факторным), то регрессия называется *парной*, если между тремя и более признаками – *множественной* (*многофакторной*) регрессией. Например, Кейнсом было предложено уравнение парной линейной регрессии, выражающей зависимость частного потребления C от располагаемого дохода Y_d: $C = C_0 + bY_d$, где $C_0 > 0$ – величина автономного потребления; $0 < b < 1$ – предельная склонность к потреблению.

При изучении регрессии следует придерживаться определенной последовательности этапов:

1. Задание аналитической формы уравнения регрессии и определение параметров регрессии.

2. Определение в регрессии степени стохастической взаимосвязи результативного признака и факторов, проверка общего качества уравнения регрессии.

3. Проверка статистической значимости каждого коэффициента уравнения регрессии и определение их доверительных интервалов.

Основное содержание выделенных этапов рассмотрим на примере множественной линейной регрессии, реализованной в режиме «Регрессия» надстройки ***Пакет анализа*** Microsoft Excel.

Этап 1. Уравнение линейной множественной регрессии имеет вид

$$\hat{y} = a_0 + a_1x_1 + a_2x_2 + ... + a_mx_m,$$

где $\hat{y}$ – теоретические значения результативного признака, полученные путем подстановки соответствующих значений факторных признаков в уравнение регрессии;

$x_1, x_2, ..., x_m$ – значения факторных признаков;

$a_0, a_1, ..., a_m$ – параметры уравнения (коэффициенты регрессии).

Параметры уравнения регрессии могут быть определены с помощью *метода наименьших квадратов** (именно этот метод и используется в Microsoft Excel). Сущность данного метода заключается в нахождении параметров модели (a_i), при которых минимизируется сумма квадратов отклонений эмпирических (фактических) значений результативного признака от теоретических, полученных по выбранному уравнению регрессии, т. е.

$$S = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{n=1}^{n}(y_i - a_0 - a_1x_{1_i} - a_2x_{2_i} - ... - a_mx_{m_i})^2 \to \min.$$

Рассматривая S в качестве функции параметров a_i и проводя математические преобразования (дифференцирование), получаем систему нормальных уравнений с m неизвестными (по числу параметров a_i):

$$\begin{cases} na_0 + a_1\sum x_1 + a_2\sum x_2 + ... + a_m\sum x_m = \sum y; \\ a_0\sum x_1 + a_1\sum x_1^2 + a_2\sum x_2x_1 + ... + a_m\sum x_mx_1 = \sum yx_1; \\ \cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots \\ a_0\sum x_m + a_1\sum x_1x_m + a_2\sum x_2x_m + ... + a_m\sum x_m^2 = \sum yx_m, \end{cases}$$

где n – число наблюдений;

m – число факторов в уравнении регрессии.

Решив систему уравнений, находим значения параметров a_i, являющихся коэффициентами искомого теоретического уравнения регрессии.

Этап 2. Для определения величины степени стохастической взаимосвязи результативного признака Y и факторов X необходимо знать следующие дисперсии:

- *общую дисперсию* результативного признака Y, отображающую влияние как основных, так и остаточных факторов:

*В справочных системах «англоязычных» программ этот метод обозначается как *LS (Least Squares Method)*.

$$\sigma_y^2 = \frac{\sum_{i=1}^{n}(y_i - \overline{y})^2}{n},$$

где $\overline{y}$ – среднее значение результативного признака Y;

• *факторную дисперсию* результативного признака Y, отображающую влияние только основных факторов:

$$\sigma_\Phi^2 = \frac{\sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2}{n};$$

• *остаточную дисперсию* результативного признака Y, отображающую влияние только остаточных факторов:

$$\sigma_O^2 = \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(m+1)}.$$

При корреляционной связи результативного признака и факторов выполняется соотношение

$$\sigma_\Phi^2 < \sigma_y^2, \text{ при этом } \sigma_y^2 = \sigma_\Phi^2 + \sigma_O^2.$$

Для анализа общего качества уравнения линейной многофакторной регрессии используют обычно *множественный коэффициент детерминации* R^2, называемый также квадратом *коэффициента множественной корреляции* R. Множественный коэффициент детерминации рассчитывается по формуле

$$R^2 = \frac{\sigma_\Phi^2}{\sigma_y^2}$$

и определяет долю вариации результативного признака, обусловленную изменением факторных признаков, входящих в многофакторную регрессионную модель.

Так как в большинстве случаев уравнение регрессии приходится строить на основе выборочных данных, то возникает вопрос

об адекватности построенного уравнения генеральным данным. Для этого проводится проверка статистической значимости коэффициента детерминации R^2 на основе F-критерия Фишера:

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m},$$

где n – число наблюдений;

m – число факторов в уравнении регрессии.

Примечание. Если в уравнении регрессии свободный член $a_0 = 0$, то числитель $n - m - 1$ следует увеличить на 1, т.е. он будет равен $n - m$.

В математической статистике доказывается, что если гипотеза $H_0 : R^2 = 0$ выполняется, то величина F имеет F-распределение с $k = m$ и $l = n-m-1$ числом степеней свободы, т.е.

$$\frac{R^2}{1-R^2} \cdot \frac{n-m-1}{m} = F(k = m, l = n-m-1).$$

Гипотеза $H_0 : R^2 = 0$ о незначимости коэффициента детерминации R^2 отвергается, если $F_p > F_{пр,\alpha}^{кр}$.

При значениях $R^2 > 0{,}7$ считается, что вариация результативного признака Y обусловлена в основном влиянием включенных в регрессионную модель факторов X.

Для оценки адекватности уравнения регрессии часто также используют показатель *средней ошибки аппроксимации*:

$$\overline{\varepsilon} = \frac{1}{n}\sum_{i=1}^{n}\frac{|y_i - \hat{y}_i|}{y_i} \cdot 100\%.$$

Этап 3. Возможна ситуация, когда часть вычисленных коэффициентов регрессии не обладает необходимой степенью значимости, т.е. значения данных коэффициентов будут меньше их стандартной ошибки. В этом случае такие коэффициенты должны быть исключены из уравнения регрессии. Поэтому проверка адекватности построенного уравнения регрессии наряду с проверкой

значимости коэффициента детерминации R^2 включает в себя также и проверку значимости каждого коэффициента регрессии.

Значимость коэффициентов регрессии проверяется с помощью t-критерия Стьюдента:

$$t = \frac{a_i}{\sigma_{a_i}},$$

где σ_{a_i} – стандартное значение ошибки для коэффициента регрессии a_i.

В математической статистике доказывается, что если гипотеза $H_0 : a_i = 0$ выполняется, то величина t имеет распределение Стьюдента с $k = n-m-1$ числом степеней свободы, т. е.

$$\frac{a_i}{\sigma_{a_i}} = t(k = n - m - 1).$$

Гипотеза $H_0 : a_i = 0$ о незначимости коэффициента регрессии отвергается, если $|t_р| > |t_{кр}|$.

Кроме того, зная значение $t_{кр}$, можно найти границы доверительных интервалов для коэффициентов регрессии:

$$a_i^{\min} = a_i - t_{кр}\sigma_{a_i};$$

$$a_i^{\max} = a_i - t_{кр}\sigma_{a_i}.$$

При экономической интерпретации уравнения регрессии также широко используются *частные коэффициенты эластичности*, показывающие, на сколько процентов в среднем изменится значение результативного признака при изменении значения соответствующего факторного признака на 1%, и определяемые по формуле

$$Э_{X_i} = a_i \frac{\overline{x}_i}{\overline{y}},$$

где $\overline{x}_i$ – среднее значение соответствующего факторного признака;
$\overline{y}$ – среднее значение результативного признака;
a_i – коэффициент регрессии при соответствующем факторном признаке.

14.2. Справочная информация по технологии работы

Режим работы «Регрессия» служит для расчета параметров уравнения *линейной* регрессии и проверки его адекватности исследуемому процессу.

В диалоговом окне данного режима (рис. 14.1) задаются следующие параметры:

1. *Входной интервал Y* – вводится ссылка на ячейки, содержащие данные по результативному признаку. Диапазон должен состоять из одного столбца.

2. *Входной интервал X* – вводится ссылка на ячейки, содержащие факторные признаки. Максимальное число входных диапазонов (столбцов) равно 16.

3. *Метки в первой строке/Метки в первом столбце* – *см.* подразд. 1.1.2.

4. *Уровень надежности* – установите данный флажок в активное состояние, если в поле, расположенное напротив флажка, необходимо ввести уровень надежности, отличный от уровня 95 %, применяемого по умолчанию. Установленный уровень надежности используется для проверки значимости коэффициента детерминации R^2 и коэффициентов регрессии a_i.

Примечание. При неактивном флажке *Уровень надежности* в таблице параметров уравнения регрессии (*см.* табл. 14.4, 14.9) генерируются две одинаковые пары столбцов для границ доверительных интервалов.

5. *Константа-ноль* – установите данный флажок в активное состояние, если требуется, чтобы линия регрессии прошла через начало координат (т. е. $a_0 = 0$).

6. *Выходной интервал/Новый рабочий лист/Новая рабочая книга* – *см.* подразд. 1.1.2.

7. *Остатки* – установите данный флажок в активное состояние, если требуется включить в выходной диапазон столбец остатков (*см.* столбец *Остатки* в табл. 14.5).

8. *Стандартизованные остатки* – установите данный флажок в активное состояние, если требуется включить в выходной диапазон столбец стандартизованных остатков (*см.* столбец *Стандартизованные остатки* в табл. 14.5).

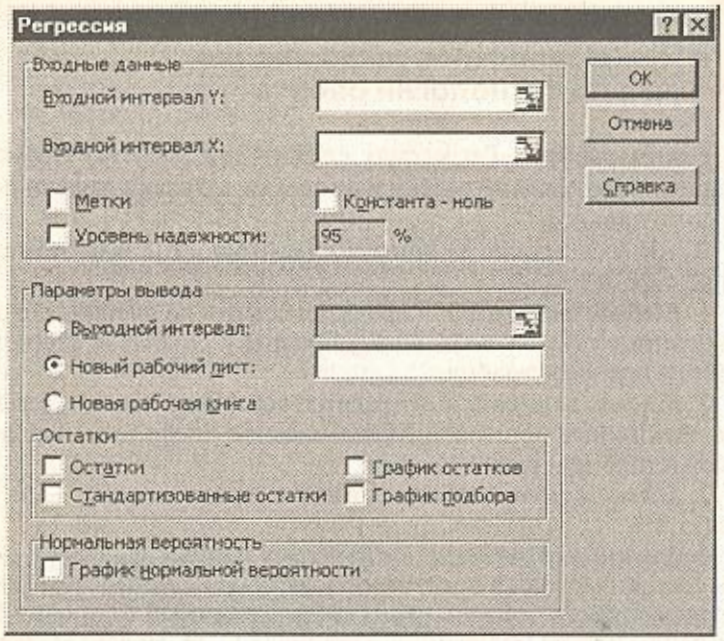

Рис. 14.1

9. *График остатков* – установите данный флажок в активное состояние, если требуется вывести на рабочий лист точечные графики зависимости остатков от факторных признаков x_i.

10. *График подбора* – установите данный флажок в активное состояние, если требуется вывести на рабочий лист точечные графики зависимости теоретических результативных значений $\hat{y}$ от факторных признаков x_i.

11. *График нормальной вероятности* – установите данный флажок в активное состояние, если требуется вывести на рабочий лист точечный график зависимости наблюдаемых значений y от автоматически формируемых интервалов персентилей. График строится на основе генерируемой таблицы «Вывод вероятности» (*см.* табл. 14.6).

Пример 14.1. Данные о прибыли предприятий Y, величине оборотных средств X_1 и стоимости основных фондов X_2 приведены в табл. 14.1, сформированной на рабочем листе Microsoft Excel.

Таблица 14.1

	B	C	D	E
2	Номер предприятия	Прибыль Y, млн руб.	Величина оборотных средств X_1, млн руб.	Стоимость основных фондов X_2, млн руб.
3	1	188	129	510
4	2	78	64	190
5	3	93	69	240
6	4	152	87	470
7	5	55	47	110
8	6	161	102	420

По представленным данным необходимо определить параметры уравнения линейной регрессии и провести его анализ.

Для решения задачи используем режим работы «Регрессия». Значения параметров, установленных в одноименном диалоговом окне, представлены на рис. 14.2, а рассчитанные в данном режиме показатели – в табл. 14.2–14.6.

Таблица 14.2

	B	C
11	ВЫВОД ИТОГОВ	
12		
13	*Регрессионная статистика*	
14	Множественный R	0,997
15	R-квадрат	0,995
16	Нормированный R-квадрат	0,991
17	Стандартная ошибка	5,050
18	Наблюдения	6

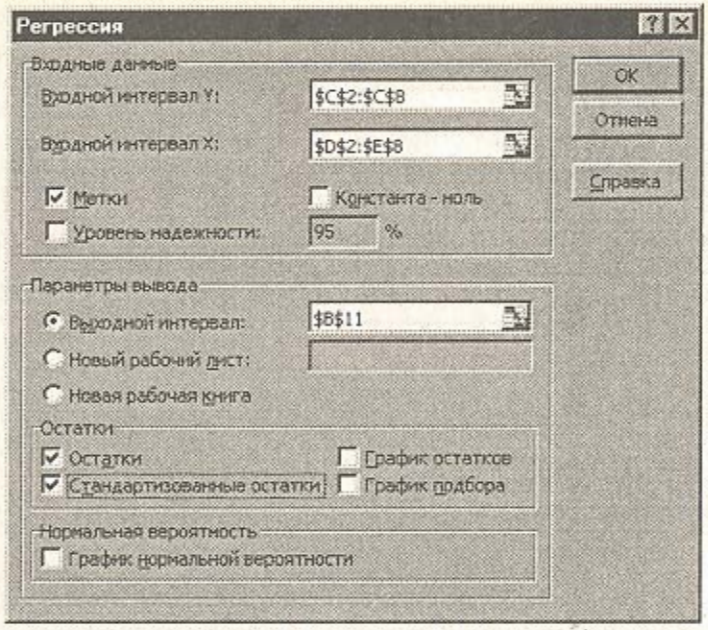

Рис. 14.2

В табл. 14.2 сгенерированы результаты по регрессионной статистике. Эти результаты соответствуют следующим статистическим показателям:

- *Множественный R* – коэффициенту корреляции R;
- *R-квадрат* – коэффициенту детерминации R^2;
- *Стандартная ошибка* – остаточному стандартному отклонению

$$\sigma_0 = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-(m+1)}};$$

- *Наблюдения* – числу наблюдений n.

В табл. 14.3 сгенерированы результаты дисперсионного анализа, которые используются для проверки значимости коэффициента детерминации R^2.

Таблица 14.3

	B	C	D	E	F	G
20	Дисперсионный анализ					
21		*df*	*SS*	*MS*	*F*	*Значимость F*
22	Регрессия	2	13962,33	6981,16	273,74	0,0004
23	Остаток	3	76,51	25,50		
24	Итого	5	14038,83			

Столбцы табл. 14.3 имеют следующую интерпретацию:

1. Столбец *df* – число степеней свободы.

Для строки *Регрессия* число степеней свободы определяется количеством факторных признаков m в уравнении регрессии $k_\Phi = m$.

Для строки *Остаток* число степеней свободы определяется числом наблюдений n и количеством переменных в уравнении регрессии $m + 1$: $k_O = n-(m+1)$.

Для строки *Итого* число степеней свободы определяется суммой $k_Y = k_\Phi + k_O$.

2. Столбец *SS* – сумма квадратов отклонений.

Для строки *Регрессия* – это сумма квадратов отклонений теоретических данных от среднего:

$$SS_\Phi^2 = \sum_{i=1}^{n}(\hat{y}_i - \overline{y})^2.$$

Для строки *Остаток* – это сумма квадратов отклонений эмпирических данных от теоретических:

$$SS_O^2 = \sum_{i=1}^{n}(y_i - \hat{y})^2.$$

Для строки *Итого* – это сумма квадратов отклонений эмпирических данных от среднего:

$$SS_Y^2 = \sum_{i=1}^{n}(y_i - \overline{y})^2 \quad \text{или} \quad SS_Y^2 = SS_\Phi^2 + SS_O^2.$$

3. Столбец *MS* – дисперсии, рассчитываемые по формуле

$$MS = \frac{SS}{df}.$$

Для строки *Регрессия* – это факторная дисперсия $\sigma^2_{Ф}$.

Для строки *Остаток* – это остаточная дисперсия $\sigma^2_{О}$.

4. Столбец *F* – расчетное значение *F*-критерия Фишера $F_р$, вычисляемое по формуле

$$F_р = \frac{MS(\text{Регрессия})}{MS(\text{Остатки})}.$$

5. Столбец *Значимость F* – значение уровня значимости, соответствующее вычисленному значению $F_р$. Определяется с помощью функции

= FРАСП($F_р$; *df*(регрессия);*df*(остаток)).

В табл. 14.4 сгенерированы значения коэффициентов регрессии a_i и их статистические оценки.

Таблица 14.4

	B	C	D	E	F	G	H	I	J
26		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
27	Y-пересечение	–1,94	7,63	–0,25	0,82	–26,21	22,32	–26,21	22,32
28	Величина оборотных средств X_1, млн руб.	0,69	0,20	3,53	0,04	0,07	1,32	0,07	1,32
29	Стоимость основных фондов X_2, млн руб.	0,20	0,04	5,75	0,01	0,09	0,31	0,09	0,31

Столбцы табл. 14.4 имеют следующую интерпретацию:

1. *Коэффициенты* – значения коэффициентов a_i.

2. *Стандартная ошибка* – стандартные ошибки коэффициентов a_i.

3. *t-статистика* – расчетные значения *t*-критерия, вычисляемые по формуле

$$t\text{-статистика} = \frac{\text{Коэффициенты}}{\text{Стандартная ошибка}}.$$

5. *P-значение* – значения уровней значимости, соответствующие вычисленным значениям $t_р$. Определяются с помощью функции

=СТЬЮДРАСП($t_р$;n–m–1).

6. *Нижние 95 %* и *Верхние 95 %* – соответственно нижние и верхние границы доверительных интервалов для коэффициентов регрессии a_i. Для нахождения границ доверительных интервалов с помощью функции = СТЬДРАСПОБР (α; $n - m - 1$) рассчитывается критическое значение *t*-критерия $t_{кр}$, а затем по формулам

Нижние 95% = Коэффициент – Стандартная ошибка · $t_{кр}$;

Верхние 95% = Коэффициент + Стандартная ошибка · $t_{кр}$

вычисляются соответственно нижние и верхние границы доверительных интервалов.

Таблица 14.5

	B	C	D	E
33	ВЫВОД ОСТАТКА			
34				
35	*Наблюдение*	*Предсказанная прибыль Y, млн руб.*	*Остатки*	*Стандартные остатки*
36	1	190,91	–2,91	–0,74
37	2	80,98	–2,98	–0,76
38	3	94,57	–1,57	–0,40
39	4	153,62	–1,62	–0,42
40	5	52,98	2,02	0,52
41	6	153,93	7,07	1,81

В табл. 14.5 сгенерированы теоретические значения $\hat{y}_i$ результативного признака Y и значения остатков. Последние вычисляются как разность между эмпирическими y и теоретическими $\hat{y}_i$ значениями результативного признака Y.

Таблица 14.6

	G	H
33	ВЫВОД ВЕРОЯТНОСТИ	
34		
35	*Персентиль*	*Прибыль Y, млн руб.*
36	8,33	55
37	25	78
38	41,67	93
39	58,33	152
40	75	161
41	91,67	188

В табл. 14.6 сгенерированы интервалы персентилей и соответствующие им эмпирические значения y.

Перейдем к анализу сгенерированных таблиц.

Рассчитанные в табл. 14.4 (ячейки C27:C29) коэффициенты регрессии a_i позволяют построить уравнение, выражающее зависимость прибыли предприятий Y от величины оборотных средств X_1 и стоимости основных фондов X_2:

$$\hat{y} = -1{,}94 + 0{,}69x_1 + 0{,}20x_2.$$

Значение множественного коэффициента детерминации $R^2 = 0{,}995$ (ячейка C15 в табл. 14.2) показывает, что 99,5 % общей вариации результативного признака объясняется вариацией факторных признаков X_1 и X_2. Значит, выбранные факторы существенно влияют на прибыль предприятий, что подтверждает правильность их включения в построенную модель.

Рассчитанный уровень значимости $\alpha_p = 0{,}0004 < 0{,}05$ (показатель *Значимость F* в табл. 14.3) подтверждает значимость R^2.

Другой подход к проверке значимости R^2 (как это делалось во всех ранее рассмотренных режимах надстройки «Пакет анализа») основан на проверке попадания Fр (показатель F в табл. 14.3) в критическую область ($F^{кр}_{пр,\,\alpha}$, $+\infty$). Для рассматриваемого примера $F^{кр}_{пр,\,\alpha} = 9{,}55$, которое рассчитывается по формуле

=FРАСПОБР(0,05;C22;C23),

где в ячейке C22 вычисляется число степеней свободы $k_Ф = m = 2$, а в ячейке C23 – число степеней свободы $k_O = n-(m+1) = 6-(2+1) = 3$.

Так как $F_р = 273{,}74$ попадает в критический интервал (9,55; $+\infty$), то гипотеза $H_0 : R^2 = 0$ отвергается, т. е. коэффициент детерминации R^2 является значимым.

Показатель средней ошибки аппроксимации $\bar{\varepsilon} = 2{,}7$ % также подтверждает достаточно высокую адекватность построенного уравнения. Данный показатель может быть рассчитан по формуле

{=СУММ(ABS(D36:D41)/(C3:C8))/СЧЕТ(C3:C8)*100},

где в массиве D36 : D41 табл. 14.5 рассчитаны разности между эмпирическими и теоретическими значениями результативного признака.

Следующим этапом является проверка значимости коэффициентов регрессии: a_0, a_1 и a_2. Сравнивая попарно элементы массивов C27:C29 и D27:D29 (*см.* табл. 14.4), видим, что абсолютное значение свободного члена a_0 меньше, чем его стандартная ошибка. Таким образом, свободный член a_0 следует исключить из уравнения регрессии.

Стандартные ошибки коэффициентов a_1 и a_2 меньше своих стандартных ошибок. К тому же эти коэффициенты являются значимыми, о чем можно судить по значениям показателя *P-значение* в табл. 14.4, которые меньше заданного уровня значимости $\alpha = 0{,}05$.

Другой распространенный способ проверки значимости коэффициентов регрессии основан на проверке попадания $t_р$ (показатель *t-статистика* в табл. 14.4) в критическую область $(-\infty, t^{кр}_{лев,\,\alpha/2}) \cup (t^{кр}_{пр,\,\alpha/2}, +\infty)$. В генерируемых таблицах режима не приводится значение $t_{кр}$, но его можно легко вычислить с помощью функции СТЬЮДРАСПОБР. Для рассматриваемого примера значение $|t_{кр}| = 3{,}18$, которое рассчитывается по формуле

=СТЬЮДРАСПОБР(0,05;6–2–1),

где 0,05 – заданный уровень значимости;
6 – число наблюдений;
2 – число факторов в уравнении регрессии;
1 – число свободных членов в уравнении регрессии.

Так как $t_p^{a_1} = 3{,}53$ и $t_p^{a_2} = 5{,}75$ попадают в критический интервал $(-\infty; -3{,}18) \cup (3{,}18; +\infty)$, то коэффициенты регрессии a_1 и a_2 являются значимыми.

Подводя итог предварительному анализу уравнения регрессии, можно сделать вывод, что его целесообразно пересчитать без свободного члена a_0, который не является статистически значимым.

Для пересчета уравнения регрессии в диалоговом окне **Регрессия** необходимо задать те же самые параметры (*см.* рис. 14.2), за исключением лишь того, что следует активизировать флажок *Константа-ноль*. В случае если незначимым является коэффициент при факторном признаке, следует пересмотреть набор признаков в уравнении регрессии.

После пересчета уравнения на рабочем листе генерируются таблицы, аналогичные табл. 14.2–14.6. Для сравнения приведем только первые три из них (табл. 14.7–14.9).

Таблица 14.7

	B	C
11	ВЫВОД ИТОГОВ	
12		
13	*Регрессионная статистика*	
14	Множественный R	0,997
15	R-квадрат	0,994
16	Нормированный R-квадрат	0,743
17	Стандартная ошибка	4,421
18	Наблюдения	6

Таким образом, получаем новое уравнение регрессии:

$$\hat{y} = 0{,}66x_1 + 0{,}21x_2.$$

Проверка значимости коэффициента детерминации R^2 и коэффициентов a_1 и a_2 при факторных признаках подтверждает адекватность полученного уравнения.

Экономическая сущность коэффициентов a_1 и a_2 в полученном уравнении регрессии состоит в том, что они показывают степень влияния каждого фактора на прибыль предприятий. Так,

Таблица 14.8

	B	C	D	E	F	G
20	Дисперсионный анализ					
21		*df*	*SS*	*MS*	*F*	*Значимость F*
22	Регрессия	2	13960,67	6980,33	357,21	0,0003
23	Остаток	4	78,16	19,54		
24	Итого	6	14038,83			

Таблица 14.9

	B	C	D	E	F	G	H	I	J
26		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
27	Y-пересечение	0	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д	#Н/Д
28	Величина оборотных средств X_1, млн руб.	0,66	0,11	5,95	0,004	0,35	0,96	0,35	0,96
29	Стоимость основных фондов X_2, млн руб.	0,21	0,03	7,65	0,002	0,13	0,28	0,13	0,28

увеличение оборотных средств на 1 млн руб. ведет к росту прибыли на 0,66 млн руб., а увеличение основных фондов на 1 млн руб. ведет к росту прибыли на 0,21 млн руб.

Кроме того, дополнительно можно рассчитать и коэффициенты эластичности $Э_{X_1} = 0{,}45$ и $Э_{X_2} = 0{,}55$, которые показывают, что по абсолютному приросту наибольшее влияние на прибыль предприятий оказывает второй фактор: увеличение стоимости основных фондов X_2 на 1 % вызывает рост прибыли на 0,55 %, тогда как рост величины оборотных средств X_1 на 1 % способствует росту прибыли на 0,45 %.

14.3. Статистические функции, связанные с режимом «Регрессия»

Функция ЛИНЕЙН

См. также ЛГРФПРИБЛ, ПРЕДСКАЗ, ТЕНДЕНЦИЯ.

Синтаксис:

ЛИНЕЙН (известные __ значения __ *y*; известные __ значения __ *x*; конст; статистика).

Результат:

Рассчитывает массив данных, описывающих уравнение линейной множественной (или парной) регрессии на основе метода наименьших квадратов.

Аргументы:

- *известные __ значения __ y*: множество значений результативного признака *Y*;
- *известные __ значения __ x*: множество значений факторных признаков X_i (необязательный аргумент);
- *конст*: логическое значение, которое указывает, требуется ли, чтобы свободный член a_0 был равен 0 (необязательный аргумент);
- *статистика*: логическое значение, которое указывает, требуется ли вернуть дополнительную статистику по регрессии (необязательный аргумент).

Замечания:

- массив *известные __ значения __ x* может содержать одно или несколько множеств переменных. Если используется только одна переменная, то аргументы *известные __ значения __ y* и *известные __ значения x* могут быть массивами любой формы при условии, что они имеют одинаковую размерность. Если используется более одной переменной, то аргумент *известные __ значения __ y* должен быть вектором (т. е. интервалом высотой в одну строку или шириной в один столбец).
- если аргумент *известные __ значения __ x* опущен, то предполагается, что это массив {1; 2; 3; ...} такого же размера, как и аргумент *известные __ значения __ y*;
- если аргумент *конст* = 1 или опущен, то a_0 вычисляется обычным образом;
- если аргумент *конст* = 0, то a_0 полагается равным 0 и значения a_i подбираются так, чтобы выполнялось соотношение $\hat{y} = a_1x_1 + a_2x_2 + ... + a_mx_m$;
- если аргумент *статистика* = 0 или опущен, то функция ЛИНЕЙН вычисляет только коэффициенты a_i и свободный член a_0;
- если аргумент *статистика* = 1, то функция ЛИНЕЙН рассчитывает дополнительную регрессионную статистику, так что возвращаемый массив будет иметь вид

$$\{a_m; a_{m-1}; ...; a_1; a_0; se_m; se_{m-1}; ...; se_1; se_0; R^2; se_y; F; df; ss_{reg}; ss_{resid}\}.$$

Дополнительная регрессионная статистика:

$se_1, se_2, ..., se_m$ – стандартные значения ошибок для коэффициентов $a_1, a_2, ..., a_m$;

se_0 – стандартное значение ошибки для свободного члена a_0 (se_0 = #Н/Д, если аргумент *конст* = 0);

R^2 – коэффициент детерминации;

se_y – стандартная ошибка для оценки *y*;

F – *F*-статистика;

df – степени свободы;

ss_{reg} – регрессионная сумма квадратов;

ss_{resid} – остаточная сумма квадратов.

Порядок расчета дополнительной регрессионной статистики представлен в табл. 14.10.

Математико-статистическая интерпретация:

См. подразд. 14.1.

Технологию работы с функцией ЛИНЕЙН рассмотрим на примере 14.1. Для данного примера формула {=ЛИНЕЙН(C3:C8;D3:E8;0;1)} рассчитает следующий массив значений (табл. 14.11):

Как видим, данная формула рассчитывает значения, аналогичные значениям из табл. 14.7–14.9:

Таблица 14.10

a_m	a_{m-1}	...	a_2	a_1	a_0
se_m	se_{m-1}	...	se_2	se_1	se_0
R^2	se_y				
F	df				
ss_{reg}	ss_{resid}				

Таблица 14.11

	E	F	G
11	0,21	0,66	0
12	0,03	0,11	#Н/Д
13	0,99	4,42	#Н/Д
14	357,21	4	#Н/Д
15	13960,67	78,16	#Н/Д

- E11 = C29 в табл. 14.9 – коэффициент a_2;
- F11 = C28 в табл. 14.9 – коэффициент a_1;
- G11 = C27 в табл. 14.9 – коэффициент a_0;
- E12 = D29 в табл. 14.9 – стандартную ошибку для коэффициента a_2;
- F12 = D28 в табл. 14.9 – стандартную ошибку для коэффициента a_1;
- E13 = C15 в табл. 14.7 – коэффициент детерминации R^2;
- F13 = C17 в табл. 14.7 – стандартную ошибку для оценки y;
- E14 = F22 в табл. 14.8 – расчетное значение F-критерия Фишера F_p;
- F14 = C23 в табл. 14.8 – число степеней свободы k_O;
- E15 = D22 в табл. 14.8 – регрессионную (факторную) сумму квадратов;
- F15 = D23 в табл. 14.8 – остаточную сумму квадратов.

Функцию ЛИНЕЙН удобно применять, когда не требуется проводить полный анализ уравнения регрессии.

Функция ТЕНДЕНЦИЯ

См. также ЛИНЕЙН, ПРЕДСКАЗ.

Синтаксис:

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

Результат:

Рассчитывает массив прогнозируемых значений результативного признака в соответствии с линейным трендом.

Аргументы:

- *известные __ значения __ y*: множество значений результативного признака *Y*;
- *известные __ значения __ x*: множество значений факторных признаков X_i (необязательный аргумент);
- *новые __ значения __ x*: множество новых значений *x*, для которых функция ТЕНДЕНЦИЯ рассчитывает соответствующие значения $\hat{y}$ (необязательный аргумент);
- *конст*: логическое значение, которое указывает, требуется ли, чтобы свободный член a_0 был равен 0 (необязательный аргумент).

Замечания:

- массив *известные __ значения __ x* может содержать одно или несколько множеств переменных. Если используется только одна переменная, то аргументы *известные __ значения __ y* и *известные __ значения __ x* могут быть массивами любой формы при условии, что они имеют одинаковую размерность. Если используется более одной переменной, то аргумент *известные __ значения __ y* должен быть вектором (т.е. интервалом высотой в одну строку или шириной в один столбец);
- если аргумент *известные __ значения __ x* опущен, то предполагается, что это массив {1;2;3;...} такого же размера, как и аргумент *известные __ значения __ y*;
- аргумент *новые __ значения __ x* должен содержать столбец (или строку) для каждой независимой переменной, так же как аргумент *известные __ значения __ x*;
- если аргумент *новые __ значения __ x* опущен, то предполагается, что он совпадает с аргументом *известные __ значения __ x*;
- если аргументы *известные __ значения __ x* и *новые __ значения __ x* опущены, то предполагается, что они являются массивами {1;2;3;...} такого же размера, что и аргумент *известные __ значения __ y*;

• если аргумент *конст* =1 или опущен, то a_0 вычисляется обычным образом;

• если аргумент *конст* =0, то a_0 полагается равным 0 и значения a_i подбираются так, чтобы выполнялось соотношение $\hat{y} = a_1x_1 + a_2x_2 + ... + a_mx_m$.

Математико-статистическая интерпретация:

Функция ТЕНДЕНЦИЯ аппроксимирует прямой линией (по методу наименьших квадратов) массивы *известные __ значения __ y* и *известные __ значения __ x* и рассчитывает в соответствии с линейным трендом новые значения $\hat{y}$ для заданного массива *новые __ значения __ x*.

Технологию работы с функцией ТЕНДЕНЦИЯ рассмотрим на примере 14.1. В этом примере было получено уравнение двухфакторной линейной регрессии $\hat{y} = 0{,}66x_1 + 0{,}21x_2$, которое позволяет получить следующие теоретические значения прибыли предприятий: 190,08; 81,28; 94,90; 154,25; 53,59; 153,76 (табл. 14.12).

Таблица 14.12

	B	C	D	E
33	ВЫВОД ОСТАТКА			
34				
35	*Наблюдение*	*Предсказанная прибыль Y, млн руб.*	*Остатки*	*Стандартные остатки*
36	1	190,08	–2,08	–0,58
37	2	81,28	–3,28	–0,91
38	3	94,90	–1,90	–0,53
39	4	154,25	–2,25	–0,62
40	5	53,59	1,41	0,39
41	6	153,76	7,24	2,01

Указанный ряд значений может быть получен и с помощью функции ТЕНДЕНЦИЯ, которая должна быть введена как формула массива для данных, приведенных в табл. 14.1: {=ТЕНДЕНЦИЯ(C3:C8;D3:E8;D3:E8;0)}.

Кроме того, если известны величина оборотных средств и стоимость основных фондов для новых предприятий, то с помощью функции ТЕНДЕНЦИЯ может быть спрогнозирована их прибыль. Например, известно, что для предприятия 7 $x_1 = 95$, $x_2 = 380$, тогда формула =ТЕНДЕНЦИЯ(C3:C8;D3:E8;{95;380}; 0) рассчитает прогнозируемое значение прибыли $\hat{y} = 140{,}90$ млн руб.

Примечание. Для парной регрессии и размерности аргумента *новые __ значения __ x* в одну ячейку функция ТЕНДЕНЦИЯ адекватна функции ПРЕДСКАЗ (*см.* описание функции ПРЕДСКАЗ).

Кроме того, функцию ТЕНДЕНЦИЯ удобно использовать при *экстраполяции* и *интерполяции рядов динамики*.

Под *экстраполяцией* понимается распространение выявленных в анализе рядов динамики закономерностей развития изучаемого явления на будущее. Экстраполяция широко применяется при прогнозировании социально-экономических явлений и базируется на следующих предпосылках:

• развитие исследуемого явления в целом следует описывать плавной кривой;

• общая тенденция развития явления в прошлом и настоящем не должна претерпевать серьезных изменений в будущем.

К прогнозированию уровней динамического ряда близок вопрос об *интерполяции* – определении некоторых неизвестных уровней внутри данного динамического ряда. Интерполяция тесно связана с аналитическим выравниванием ряда (*см.* главу 16). При интерполяции считается, что ни выявленная тенденция, ни ее характер не претерпели существенных изменений в том промежутке времени, уровни которого нам известны. Такое предположение обычно является более обоснованным, чем предположение о будущей тенденции.

Функцию ТЕНДЕНЦИЯ можно также использовать и для аппроксимации полиномиальной кривой $\hat{y} = a_0 + a_1x + a_2x^2 + ... + a_mx^m$, проводя регрессионный анализ для одного факторного признака, но возведенного в различные степени. Например, пусть столбец B содержит значения y, а столбец C – значения x. Можно ввести x^2 в столбец D, x^3 – в столбец E и так далее, а затем на основании введенных данных построить уравнение регрессии. Например, большое применение при выравнивании ря-

дов динамики имеет парабола (полином 2-го порядка), уравнение которой имеет вид

$$\hat{y} = a_0 + a_1 t + a_2 t^2.$$

Выбор параболы основывается на предположении о том, что не скорость, а ускорение является постоянной величиной. В переводе на язык экономики это будет означать предположение, что абсолютные приросты данного ряда динамики не стабильны, а обнаруживают тенденцию к изменению на некоторую постоянную величину. Так, например, если было бы установлено, что ежегодно абсолютный прирост урожайности в среднем увеличивается на определенное количество центнеров с гектара, то в таком случае выравнивание ряда нужно было бы производить по параболе 2-го порядка.

Пример 14.2. Известны данные об урожайности пшеницы в области за 1993–1999 гг. Какую урожайность пшеницы можно ожидать в области в 2000–2002 гг.?

Год	1993	1994	1995	1996	1997	1998	1999
Урожайность, ц/га	25,0	25,3	25,7	26,2	26,9	27,8	28,7

Анализ представленных данных позволяет заметить, что в рассматриваемом периоде рост урожайности пшеницы происходил с некоторым ускорением (можно показать, что абсолютный прирост урожайности в среднем увеличивался на 0,12 ц в год). Учитывая данное обстоятельство, проведем выравнивание ряда по параболе (табл. 14.13).

Содержимое ячеек в табл. 14.13:

- массив B3:C9 содержит исходные данные задачи;
- массив D3:D9 содержит года прошедшего периода, а массив E3:E9 – их квадраты (например, ячейка E3 содержит формулу =СТЕПЕНЬ(D3;2);

Таблица 14.13

	B	C	D	E	F
2	Год	Фактическая урожайность y, ц/га	x	x^2	Теоретическая урожайность $\hat{y}$, ц/га
3	1993	25,0	1993	3972049	25,01
4	1994	25,3	1994	3976036	25,28
5	1995	25,7	1995	3980025	25,69
6	1996	26,2	1996	3984016	26,23
7	1997	26,9	1997	3988009	26,92
8	1998	27,8	1998	3992004	27,75
9	1999	28,7	1999	3996001	28,72
10	**2000**		2000	4000000	**29,83**
11	**2001**		2001	4004001	**31,08**
12	**2002**		2002	4008004	**32,47**

- массив F3:F9 содержит формулу {=ТЕНДЕНЦИЯ(C3:C9; D3:E9;D3:E9;1)} – рассчитывается массив значений теоретического ряда урожайности $\hat{y}$;

Примечание. Для ввода формулы необходимо предварительно выделить диапазон ячеек F3:F9, после чего ввести формулу и нажать комбинацию клавиш **Ctrl + Shift + Enter**. Microsoft Excel автоматически заключит формулу в фигурные скобки {}.

- массив D10:D12 содержит порядковые номера годов в прогнозируемом периоде, а массив E10:E12 – их квадраты (например, ячейка E10 содержит формулу =СТЕПЕНЬ(D10;2);
- массив F10:F12 содержит формулу {=ТЕНДЕНЦИЯ(C3:C9; D3:E9; D10:E12;1)} – вычисляется массив прогнозируемых значений урожайности на 2000–2002 гг.

Таким образом, при сохранении тенденции, которая наблюдалась в течение последних семи лет, можно ожидать, что урожайность пшеницы в области в последующие три года составит приблизительно 29,8; 31,1; 32,5 ц/га.

Функция ПРЕДСКАЗ

См. также ЛИНЕЙН, ТЕНДЕНЦИЯ.

Синтаксис:

ПРЕДСКАЗ (*x*; известные __ значения __ *y*; известные __ значения __ *x*)

Результат:

Рассчитывает для *парной* регрессии прогнозируемое значение результативного признака в соответствии с *линейным* трендом.

Аргументы:

- *x*: точка данных, для которой предсказывается значение;
- *известные __ значения __ y*: множество значений результативного признака *Y*;
- *известные __ значения __ x*: множество значений факторного признака *X*.

Замечания:

- если аргумент *x* не является числом, то функция ПРЕДСКАЗ помещает в ячейку значение ошибки #ЗНАЧ!;
- если аргументы *известные __ значения __ y* и *известные __ значения __ x* пусты или содержат различное количество точек данных, то функция ПРЕДСКАЗ помещает в ячейку значение ошибки #Н/Д;
- если дисперсия аргумента *известные __ значения __ x* равна 0, то функция ПРЕДСКАЗ помещает в ячейку значение ошибки #ДЕЛ/0!.

Математико-статистическая интерпретация:

См. описание функции ТЕНДЕНЦИЯ.

Функция ПРЕДСКАЗ является частным случаем функции ТЕНДЕНЦИЯ, когда последняя применяется к *парной* регрессии и ее аргумент *новые __ значения __ x* имеет размерность в одну ячейку.

Пример 14.3. Покупатель планирует приобрести квартиру в декабре текущего года. В июне он собирает информацию о ценах на подобную квартиру за последние 6 мес. Какую цену может ожидать покупатель в декабре?

Рассмотрим решение задачи в среде Microsoft Excel (табл. 14.14).

Содержимое ячеек в табл. 14.14:

- массив B3:C8 содержит исходные данные задачи;
- массив D3:D8 содержит порядковые номера месяцев в рассматриваемом периоде;

Таблица 14.14

	B	C	D
2	Месяц	Стоимость квартиры, у. е.	Порядковый номер месяца
3	январь	22500	1
4	февраль	22600	2
5	март	22750	3
6	апрель	22700	4
7	май	22780	5
8	июнь	22800	6
9	Прогноз на декабрь	**23172**	**12**

- ячейка C9 содержит формулу =ПРЕДСКАЗ(12;C3:C8;D3:D8) – рассчитывается прогнозируемая стоимость квартиры в декабре (12 – порядковый номер декабря).

Таким образом, при сохранении тенденции, которая наблюдалась в течение последних шести месяцев, можно ожидать, что стоимость квартиры в декабре текущего года составит приблизительно 23172 у. е.

Примечание. Аналогичное решение может быть получено и с помощью функции ТЕНДЕНЦИЯ. Для этого в ячейку C9 необходимо ввести формулу =ТЕНДЕНЦИЯ(C3:C8;D3:D8;12;1).

Функция НАКЛОН

См. также ЛИНЕЙН.

Синтаксис:

НАКЛОН (известные __ значения __ *y*; известные __ значения __ *x*)

Результат:

Рассчитывает наклон прямой линии для *парной линейной* регрессии.

Аргументы:

- *известные __ значения __ y*: множество значений результативного признака *Y*;
- *известные __ значения __ x*: множество значений факторного признака *X*.

Замечания:

• аргументы должны быть числами или именами, массивами или ссылками, содержащими числа;

• если аргумент, который является массивом или ссылкой, содержит текстовые, логические значения или пустые ячейки, то такие значения игнорируются; однако ячейки с нулевыми значениями учитываются;

• если аргументы *известные __ значения __ y* и *известные __ значения __ x* пусты или содержат различное число точек данных, то функция СТОШУХ помещает в ячейку значение ошибки #Н/Д.

Математико-статистическая интерпретация:

Наклон линии линейной регрессии является мерой скорости изменения результативного признака *Y* относительно изменения факторного признака *X* и определяется отношением расстояния по вертикали на расстояние по горизонтали между двумя любыми точками теоретической прямой:

$$a_1 = \frac{\hat{y}_j - \hat{y}_i}{x_j - x_i}, \; i \neq j.$$

В уравнении парной регрессии наклон линии определяется коэффициентом a_1 и показывает, насколько изменится в среднем значение результативного признака при увеличении факторного на единицу собственного измерения.

Значение наклона линии регрессии (коэффициента a_1) удобно находить с помощью функции НАКЛОН, исключающей предварительные расчеты. Заметим, что это же значение рассчитывает и функция ЛИНЕЙН в таблице дополнительной регрессионной статистики (*см.* значение a_1 в табл. 14.10).

Пример 14.4. Требуется рассчитать наклон линии линейной регрессии для примера 14.3.

Формула =НАКЛОН(C3:C8;D3:D8) рассчитает искомое значение a_1 = 56,86. Это же значение вычисляет и формула =ЛИНЕЙН(C3:C8;D3:D8;1;1) в таблице дополнительной регрессионной статистики (*см.* значение a_1 в табл. 14.10). Кроме того, аналогичное решение можно получить с помощью формулы

$$a_1 = \frac{\sum_{i=1}^{n}(y_i - \bar{y})(x_i - \bar{x})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}.$$

Вместе с тем заметим, что данная формула требует довольно громоздких расчетов, поэтому более предпочтительным является использование функции НАКЛОН.

Функция ОТРЕЗОК

См. также ЛИНЕЙН.

Синтаксис:

ОТРЕЗОК (известные __ значения __ *x*; известные __ значения __ *y*)

Результат:

Рассчитывает значение, соответствующее точке пересечения линии *парной линейной* регрессии с осью *Y*.

Аргументы:

• *известные __ значения __ y*: множество значений результативного признака *Y*;

• *известные __ значения __ x*: множество значений факторного признака *X*.

Замечания:

• аргументы должны быть числами или массивами, содержащими числа;

• если аргумент, который является массивом, содержит текстовые, логические значения или пустые ячейки, то такие значения игнорируются; однако ячейки с нулевыми значениями учитываются;

• если аргументы *известные __ значения __ y* и *известные __ значения __ x* пусты или содержат различное число точек данных, то функция ОТРЕЗОК помещает в ячейку значение ошибки #Н/Д.

Математико-статистическая интерпретация:

Функция ОТРЕЗОК используется, когда нужно определить значение результативного признака при значении факторного признака, равном 0. В этом случае уравнение $\hat{y} = a_0 + a_1 x$ принимает вид $\hat{y} = a_0$.

Например, функцию ОТРЕЗОК можно использовать, чтобы предсказать электрическое сопротивление металла при темпера-

туре 0°С, если имеются данные измерений при комнатной температуре и выше.

Пример 14.5. Для задачи, приведенной в примере 14.3, рассчитать, какую предположительно стоимость имела квартира в декабре прошедшего года.

В этой задаче линия регрессии строится на основании данных за первые шесть месяцев текущего года, где 1 – январь. Поэтому для построенной линии регрессии декабрь предыдущего года будет иметь значение 0, ноябрь – значение –1, октябрь значение –2 и т.д. Тогда формула =ОТРЕЗОК(C3:C8;D3:D8) рассчитает искомое значение стоимости квартиры за декабрь предыдущего года, равное 22489,33 у. е.

Заметим, что это же значение вычисляет и функция ЛИНЕЙН в таблице дополнительной регрессионной статистики (*см.* значение a_0 в табл. 14.10). Кроме того, аналогичное решение можно получить и с помощью формулы =ТЕНДЕНЦИЯ(C3:C8;D3:D8;0;1) (здесь 0 – значение факторного признака *X*, для которого функция ТЕНДЕНЦИЯ рассчитывает соответствующее значение результативного признака *Y*).

Функция СТОШYX

См. также ЛИНЕЙН.

Синтаксис:

СТОШYX (известные __ значения __ *y*; известные __ значения __ *x*)

Результат:

Рассчитывает для *парной линейной* регрессии стандартную ошибку оценки результативного признака *Y*.

Аргументы:

• *известные __ значения __ y*: множество значений результативного признака *Y*;

• *известные __ значения __ x*: множество значений факторного признака *X*.

Замечания:

• аргументы должны быть числами или массивами, содержащими числа;

• если аргумент, который является массивом или ссылкой, содержит текстовые, логические значения или пустые ячейки, то такие значения игнорируются; однако ячейки с нулевыми значениями учитываются;

• если аргументы *известные __ значения __ y* и *известные __ значения __ x* пусты или содержат различное число точек данных, то функция СТОШYX помещает в ячейку значение ошибки #Н/Д.

Математико-статистическая интерпретация:

Стандартная ошибка оценки результативного признака *Y* (далее просто *стандартная ошибка оценки)* является мерой среднего рассеивания наблюденных значений (точек) вокруг подобранной линии регрессии, тем самым давая некоторое представление о надежности уравнения регрессии для производства прогнозных расчетов. Для парной регрессии стандартная ошибка оценки определяется следующим образом:

$$\sigma_{ош} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

где y_i – *i*-е фактическое значение результативного признака;
$\hat{y}_i$ – *i*-е теоретическое значение результативного признака;
n – объем выборочной совокупности.

Для парной регрессии значение стандартной ошибки оценки удобно находить с помощью функции СТОШYX, исключающей предварительные расчеты. Заметим, что это же значение рассчитывает и функция ЛИНЕЙН в таблице дополнительной регрессионной статистики (*см.* значение se_y в табл. 14.10).

Пример 14.6. Требуется рассчитать стандартную ошибку $\sigma_{ош}$ для примера 14.3.

Формула =СТОШYX(C3:C8;D3:D8) рассчитает искомое значение $\sigma_{ош}$=53,64. Это же значение вычисляет и формула =ЛИНЕЙН(C3:C8;D3:D8;1;1) в таблице дополнительной регрессионной статистики (*см.* значение se_y в табл. 14.10).

Необходимо отметить, что функцию СТОШYX нельзя использовать применительно к множественной регрессии. Для этого необходимо использоваться функцию ЛИНЕЙН с аргументом *статистика* =1 или следующую формулу:

$$\sigma_{ош} = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-m-1}}.$$

Примечание. В множественной регрессии для получения теоретических значений результативного признака *Y* удобно использовать функцию ТЕНДЕНЦИЯ.

Функция ПИРСОН

См. также КОРРЕЛ, ЛИНЕЙН.

Синтаксис:

ПИРСОН (массив1; массив2)

Результат:

Рассчитывает значение коэффициента корреляции Пирсона для *парной линейной* регрессии (аналогично функции КОРРЕЛ).

Аргументы:

- *массив1*: множество значений факторного признака *X*;
- *массив2*: множество значений результативного признака *Y*.

Замечания:

- аргументы должны быть числами или массивами, содержащими числа;
- если аргумент, который является массивом, содержит текстовые, логические значения или пустые ячейки, то такие значения игнорируются; однако ячейки с нулевыми значениями учитываются;
- если аргументы *массив1* или *массив2* пусты или содержат различное число точек данных, то функция ПИРСОН помещает в ячейку значение ошибки #Н/Д.

Математико-статистическая интерпретация:

Функция ПИРСОН, так же как и функция КОРРЕЛ, рассчитывает значение линейного коэффициента корреляции* между двумя множествами данных.

Между линейным коэффициентом корреляции и коэффициентом регрессии существует определенная зависимость, выражаемая формулой

$$r = a_1 \frac{\sigma_\Phi}{\sigma_y},$$

* Линейный коэффициент корреляции получил также название коэффициента корреляции Пирсона.

где a_1 – коэффициент при факторном признаке в уравнении регрессии (a_1 определяет наклон линии регрессии – *см.* описание функции НАКЛОН);

σ_Φ – стандартное отклонение факторного признака;

σ_y – стандартное отклонение результативного признака.

Так, для примера 14.3 формула =ПИРСОН(D3:D8;C3:C8) рассчитывает значение 0,91, такое же, как и формула =НАКЛОН(C3:C8;D3:D8)*СТАНДОТКЛОНП(D3:D8)/СТАНДОТКЛОНП(C3:C8).

Функция КВПИРСОН

См. также КОРРЕЛ, ЛИНЕЙН, ПИРСОН.

Синтаксис:

КВПИРСОН (известные __ значения __ *y*; известные __ значения __ *x*)

Результат:

Рассчитывает квадрат коэффициента корреляции Пирсона для *парной* линейной регрессии.

Аргументы:

- *известные __ значения __ y*: множество значений результативного признака *Y*;
- *известные __ значения __ x*: множество значений факторного признака *X*.

Замечания:

- аргументы должны быть числами или массивами, содержащими числа;
- если аргумент, который является массивом, содержит текстовые, логические значения или пустые ячейки, то такие значения игнорируются; однако ячейки с нулевыми значениями учитываются;
- если аргументы *известные __ значения __ y* и *известные __ значения __ x* пусты или содержат различное число точек данных, то функция КВПИРСОН помещает в ячейку значение ошибки #Н/Д.

Математико-статистическая интерпретация:

См. описание функции ПИРСОН.

Функция КВПИРСОН рассчитывает квадрат коэффициента корреляции Пирсона для парной линейной регрессии.

Так, для примера 14.3 формула =КВПИРСОН(C3:C8;D3:D8) рассчитывает значение 0,83, такое же, как и формула =СТЕПЕНЬ(ПИРСОН(D3:D8;C3:C8);2).

14.4. Родственные статистические функции

Функция ЛГРФПРИБЛ

См. также ЛИНЕЙН, РОСТ.

Синтаксис:

ЛГРФПРИБЛ (известные__значения__*y*; известные__значения__*x*; конст; статистика).

Результат:

Рассчитывает массив данных, описывающих уравнение *экспоненциальной* регрессии.

Аргументы:

• *известные__значения__y*: множество значений результативного признака *Y*;

• *известные__значения__x*: множество значений факторных признаков X_i (необязательный аргумент);

• *конст:* логическое значение, которое указывает, требуется ли, чтобы коэффициент a_0 был равен 1 (необязательный аргумент);

• *статистика*: логическое значение, которое указывает, требуется ли вернуть дополнительную статистику по регрессии (необязательный аргумент).

Замечания:

• массив *известные__значения__x* может содержать одно или несколько множеств переменных. Если используется только одна переменная, то аргументы *известные__значения__y* и *известные__значения__x* могут быть массивами любой формы при условии, что они имеют одинаковую размерность. Если используется более одной переменной, то аргумент *известные__значения__y* должен быть вектором (т. е. интервалом высотой в одну строку или шириной в один столбец);

• если аргумент *известные__значения__x* опущен, то предполагается, что это массив {1;2;3;...} такого же размера, как и аргумент *известные__значения__y*;

• если аргумент *конст* = 1 или опущен, то a_0 вычисляется обычным образом;

• если аргумент *конст* = 0, то a_0 полагается равным 1 и значения a_i подбираются так, чтобы выполнялось соотношение $\hat{y} = a_1^{x_1} a_2^{x_2} \ldots a_m^{x_m}$;

• если аргумент *статистика* = 0 или опущен, то функция ЛГРФПРИБЛ вычисляет только коэффициенты a_i (в том числе и a_0);

• если аргумент *статистика* = 1, то функция ЛГРФПРИБЛ рассчитывает дополнительную регрессионную статистику, так что возвращаемый массив будет иметь вид: $\{a_m;a_{m-1};\ldots;a_1;a_0;se_m;se_{m-1};\ldots;se_1;se_0;R^2;se_y;F;df;ss_{reg};ss_{resid}\}$.

Дополнительная регрессионная статистика:

$se_1, se_2, \ldots, se_m$: стандартные значения ошибок для коэффициентов $a_1, a_2, \ldots, a_m$;

se_0 : стандартное значение ошибки для коэффициента a_0 (se_0 = = #Н/Д, если аргумент *конст* = 0);

R^2 : коэффициент детерминации;

se_y : стандартная ошибка для оценки *y*;

F: *F*-статистика;

df : степени свободы;

ss_{reg} : регрессионная сумма квадратов;

ss_{resid}: остаточная сумма квадратов.

Порядок расчета дополнительной регрессионной статистики представлен в табл. 14.15.

Таблица 14.15

a_m	a_{m-1}	...	a_2	a_1	a_0
se_m	se_{m-1}	...	se_2	se_1	se_0
R^2	se_y				
F	df				
ss_{reg}	ss_{resid}				

Внимание! Дополнительная регрессионная статистика, которую рассчитывает функция ЛГРФПРИБЛ, основана на следующей линейной модели:

$$\ln\hat{y} = \ln a_0 + x_1 \ln a_1 + x_2 \ln a_2 + \ldots + x_m \ln a_m.$$

Это следует помнить при оценке дополнительной регрессионной статистики. Например, для расчета значимости коэффициентов регрессии используется формула $t_p = \ln(|a_i|)/\sigma_{a_i}$ (сравните с формулой для линейной регрессии $t_p = |a_i|/\sigma_{a_i}$.

Математико-статистическая интерпретация:

См. описание функции ЛИНЕЙН.

Если прямая линия отражает закон изменений в арифметической прогрессии, то линией, отражающей закон роста в геометрической прогрессии, является *показательная (экспоненциальная)* кривая.

Уравнение показательной (экспоненциальной) множественной регрессии имеет следующий вид:

$$\hat{y} = a_0 a_1^{x_1} a_2^{x_2} \ldots a_m^{x_m},$$

где $\hat{y}$ – теоретические значения результативного признака, полученные в результате подстановки соответствующих значений факторных признаков в уравнение регрессии;

$x_1, x_2, \ldots, x_m$ – значения факторных признаков;

$a_0, a_1, \ldots, a_m$ – параметры уравнения (коэффициенты регрессии).

Выравнивание по показательной (экспоненциальной) кривой широко применяется в практике статистических исследований, поскольку характер динамики многих социально-экономических явлений (увеличение объема промышленной продукции, рост капитальных вложений, рост численности персонала в той или иной отрасли и т.д.) соответствует гипотезе о росте в геометрической прогрессии. Особенно часто выравнивание по показательной (экспоненциальной) кривой применяется для рядов динамики с равноотстоящими уровнями, в которых промежуток времени между взятыми годами составляет не один год, а несколько лет.

Техника выравнивания по показательной (экспоненциальной) кривой не отличается от техники выравнивания по прямой линии с той только существенной разницей, что выравниванию по прямой подвергаются не сами члены ряда, а их логарифмы.

Пример 14.7. Требуется по данным о прибыли предприятий Y, величине оборотных средств X_1 и стоимости основных фондов X_2 определить зависимость между результативным и факторными признаками (табл. 14.16) (сравните с похожим примером 14.1).

Содержимое ячеек в табл. 14.16:

- массив B3:E8 содержит исходные данные задачи;
- массив C10:E14 содержит формулу {=ЛГРФПРИБЛ(C3:C8; D3:E8;1;1)} – вычисляется массив значений регрессионной статистики);

Таблица 14.16

	B	C	D	E
2	Номер предприятия	Прибыль Y, млн руб.	Величина оборотных средств X_1, млн руб.	Стоимость основных фондов X_2, млн руб
3	1	352	115	510
4	2	72	59	190
5	3	86	69	230
6	4	310	87	470
7	5	52	42	110
8	6	161	135	445
9	Регрессионная статистика:			
10		1,007	0,989	39,576
11		0,0004	0,002	0,087
12		0,995	0,075	#Н/Д
13		277,233	3,000	#Н/Д
14		3,131	0,017	#Н/Д
15	Статистический анализ модели:			
16	F-статистика:		t-статистика:	
17	$F_р$ =	277,23	$t_р^{a_0}$ =	42,11
18	$F_{кр}$ =	9,55	$t_р^{a_1}$ =	5,37
19			$t_р^{a_2}$ =	15,69
20			$t_{кр}$ =	3,18

Примечание. Для ввода формулы {=ЛГРФПРИБЛ(C3:C8;D3:E8;1;1)} необходимо предварительно выделить диапазон ячеек C10 : E14, после чего ввести формулу и нажать комбинацию клавиш **Ctrl + Shift + Enter**. Microsoft Excel автоматически заключит формулу в фигурные скобки {}.

- ячейка C17 содержит формулу = C13 – находится расчетное значение F-критерия $F_р$;
- ячейка C18 содержит формулу =FРАСПОБР(0,05;2;3) – рассчитывается табличное значение F-критерия $F_{кр}$ ($\alpha = 0,05$; $k = m = 2$; $l = n - m - 1 = 6 - 2 - 1 = 3$). Выполнение неравенства

$F_р > F_{кр}$ свидетельствует об адекватности построенного уравнения регрессии исследуемому процессу;

• ячейка E17 содержит формулу =ABS(LN(E10)/E11) – определяется расчетное значение *t*-критерия для коэффициента $a_0(t_р^{a_0})$;

• ячейка E18 содержит формулу =ABS(LN(D10)/D11) – вычисляется расчетное значение *t*-критерия для коэффициента $a_1(t_р^{a_1})$;

• ячейка E19 содержит формулу =ABS(LN(C10)/C11) – находится расчетное значение *t*-критерия для коэффициента a_2 $(t_р^{a_2})$;

• ячейка E20 содержит формулу =СТЬЮДРАСПОБР(0,05; 3) – рассчитывается табличное значение *t*-критерия $t_{кр}(\alpha = 0{,}05;\ l = n-m-1=6-2-1=3)$. Выполнение неравенств $|t_р^{a_0}| > |t_{кр}|$, $|t_р^{a_1}| > |t_{кр}|$ и $|t_р^{a_2}| > |t_{кр}|$ свидетельствует о значимости коэффициентов регрессии a_0, a_1 и a_2.

Рассчитанные данные (ячейки C10:E10) позволяют построить уравнение регрессии, выражающей зависимость прибыли предприятий *Y* от величины оборотных средств X_1 и стоимости основных фондов X_2:

$$\hat{y} = 39{,}576 \cdot 0{,}989^{x_1} \cdot 1{,}007^{x_2}.$$

В построенном уравнении все коэффициенты регрессии a_0, a_1 и a_2 являются значимыми, значимым является и коэффициент детерминации $R^2 = 0{,}995$, следовательно, построенное уравнение является адекватным исследуемому процессу.

Функция РОСТ

См. также ЛГРФПРИБЛ, ТЕНДЕНЦИЯ.

Синтаксис:

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

Результат:

Рассчитывает массив прогнозируемых значений результативного признака в соответствии с *экспоненциальной* кривой.

Аргументы:

• *известные __ значения __ y*: множество значений результативного признака *Y*;

• *известные __ значения __ x*: множество значений факторных признаков X_i (необязательный аргумент);

• *новые __ значения __ x*: множество новых значений *x*, для которых функция РОСТ рассчитывает соответствующие значения $\hat{y}$ (необязательный аргумент);

• *конст*: логическое значение, которое указывает, требуется ли, чтобы коэффициент a_0 был равен 1 (необязательный аргумент).

Замечания:

• если какие-либо числа в массиве *известные __ значения __ y* равны 0 или отрицательны, то функция РОСТ помещает в ячейку значение ошибки #ЧИСЛО!;

• массив *известные __ значения __ x* может содержать одно или несколько множеств переменных. Если используется только одна переменная, то аргументы *известные __ значения __ y* и *известные __ значения __ x* могут быть массивами любой формы при условии, что они имеют одинаковую размерность. Если используется более одной переменной, то аргумент *известные __ значения __ y* должен быть вектором (т.е. интервалом высотой в одну строку или шириной в один столбец).

• если аргумент *известные __ значения __ x* опущен, то предполагается, что это массив {1;2;3;...} такого же размера, как и аргумент *известные __ значения __ y*;

• если аргумент *новые __ значения __ x* опущен, то предполагается, что он совпадает с аргументом *известные __ значения __ x*;

• если аргументы *известные __ значения __ x* и *новые __ значения __ x* опущены, то предполагается, что это массивы {1;2;3;...} такого же размера, что и аргумент *известные __ значения __ y*;

• если аргумент *конст* = 1 или опущен, то a_0 вычисляется обычным образом;

• если аргумент *конст* = 0, то a_0 полагается равным 1 и значения a_i подбираются так, чтобы выполнялось соотношение $\hat{y} = a_1^{x_1} a_2^{x_2} \ldots a_m^{x_m}$.

Математико-статистическая интерпретация:

См. описание функции ЛГРФПРИБЛ, ТЕНДЕНЦИЯ.

Функция РОСТ аппроксимирует показательной (экспоненциальной) кривой массивы *известные __ значения __ y* и *известные __ значения __ x* и рассчитывает в соответствии с этой кривой новые значения $\hat{y}$ для заданного массива *новые __ значения __ x*.

Функцию РОСТ удобно использовать при *экстраполяции* и *интерполяции* рядов динамики, для которых присуща тенденция роста в геометрической прогрессии.

Пример 14.8. В примере 14.7 было получено аналитическое выражение показательной (экспоненциальной) регрессии, которое

позволяет получить следующие теоретические значения прибыли предприятий: 319,28; 72,45; 84,45; 332,16; 51,64; 168,80. Например, для предприятия 1 значение 319,28 рассчитывается по формуле

=E10*СТЕПЕНЬ(D10;D3)*СТЕПЕНЬ(C10;E3),

где в ячейках E10, D10 и C10 (*см.* табл. 14.16) рассчитываются значения коэффициентов a_0, a_1 и a_2;
в ячейках D3 и E3 (*см.* табл. 14.16) содержатся данные по предприятию 1.

Указанный ряд значений может быть получен и с помощью функции РОСТ, которая должна быть введена как формула массива: {=РОСТ(C3:C8;D3:E8;D3:E8;1)}.

Кроме того, если известны величина оборотных средств и стоимость основных фондов для новых предприятий, то с помощью функции РОСТ может быть спрогнозирована и их прибыль. Например, известно, что для предприятия 7 $x_1 = 135$, $x_2 = 530$, тогда формула =РОСТ(C3:C8;D3:E8;{135;530};1) рассчитает прогнозируемое значение прибыли $\hat{y} = 293{,}5$ млн руб.

РАЗДЕЛ V

Статистические методы изучения динамики процессов

ГЛАВА 15
Скользящее среднее и экспоненциальное сглаживание

15.1. Краткие сведения из теории статистики

Экономические данные (со статистической точки зрения) обычно делятся на два вида: перекрестные данные (cross-section data) и временные ряды (time series) [3].

Перекрестные данные – это данные по какому-либо экономическому показателю, полученные для разных однотипных объектов (предприятий, фирм, регионов и т. п.). При этом либо все данные относятся к одному и тому же моменту времени, либо их временная принадлежность несущественна. Анализ именно таких данных и проводился в предыдущих главах*.

Временной ряд представляет собой последовательность измерений в последовательные моменты времени. В отличие от анализа перекрестных данных анализ временных рядов основывается на предположении, что последовательные значения в наборе данных наблюдаются через равные промежутки времени (тогда как в других методах не важна и часто не интересна привязка наблюдений ко времени).

Анализ временных рядов включает широкий спектр разведочных процедур и исследовательских методов, которые ставят две

*Некоторое исключение составляют функции ПРЕДСКАЗ и ТЕНДЕНЦИЯ (*см.* поразд. 14.3), в описании которых преведено несколько простых примеров анализа временных рядов.

основные цели: определение природы временного ряда и предсказание будущих значений временного ряда по настоящим и прошлым значениям (прогнозирование). Обе эти цели требуют, чтобы модель ряда была идентифицирована и более или менее формально описана.

Как и большинство других видов анализа, анализ временных рядов предполагает, что данные содержат систематическую составляющую (обычно включающую несколько компонент) и случайный шум (ошибку), который затрудняет обнаружение регулярных компонент. В зависимости от формы разложения временного ряда на систематическую d и случайную составляющие e различают *аддитивную* ($\hat{y} = d + e$) и *мультипликативную* ($\hat{y} = de$) модели временного ряда. В свою очередь, в систематической компоненте временного ряда d обычно выделяют три составляющие: *тренд* tr, *сезонную компоненту* s и *циклическую компоненту* c. Таким образом, например, аддитивную модель временного ряда можно представить следующим образом:

$$\hat{y} = tr + s + c + e.$$

В зависимости от того, изменяются или не изменяются во времени вероятностные свойства (математическое ожидание, дисперсия) изучаемой случайной величины, различают *нестационарные* и *стационарные* временные ряды. Экономические процессы обычно не являются стационарными, так как содержат систематическую составляющую, но их можно преобразовать в стационарные путем исключения тренда, сезонной и циклической компонент.

Существует достаточно большое число методов сведения ряда к стационарности. Например, для выделения тренда широкое распространение получили *метод наименьших квадратов* (принципы метода рассматривались в подразд. 14.1) и *метод простых разностных операторов*, для выделения сезонной компоненты – *метод сезонного выравнивания* и *метод сезонных разностных операторов*, для выделения тренда и циклической компоненты – *метод скользящей средней* и *метод экспоненциального сглаживания*.

Рассмотрим два последних метода более подробно.

Метод скользящей средней. Это один из самых старых и широко известных способов сглаживания временного ряда. Сглаживание представляет собой некоторый способ локального усреднения данных, при котором несистематические компоненты взаимно погашают друг друга. Так, метод скользящей средней основан на переходе от начальных значений ряда к их средним значениям на интервале времени, длина которого выбрана заранее (данный интервал времени часто называют «окном»). При этом сам выбранный интервал скользит вдоль ряда.

Получаемый таким образом ряд скользящих средних ведет себя более гладко, чем исходный ряд, за счет усреднения отклонений исходного ряда. Таким образом, эта процедура дает представление об общей тенденции поведения ряда. Ее применение особенно полезно для рядов с сезонными колебаниями и неясным характером тренда. В частности, переход к ряду скользящих средних может быть использован для выявления сезонной компоненты (или сезонного индекса) временного ряда (*см.* пример 15.1).

Применяя метод скользящей средней, вместо средней можно использовать медиану значений, попавших в окно. Основное преимущество *медианного сглаживания* в сравнении со сглаживанием скользящей средней состоит в том, что результаты становятся более устойчивыми к выбросам, имеющимся внутри окна. Основной недостаток медианного сглаживания в том, что при отсутствии явных выбросов он приводит к более «зубчатым» кривым, чем сглаживание скользящей средней, и не позволяет использовать веса.

Дадим некоторое формальное определение методу скользящей средней для окна сглаживания, длина которого выражается нечетным числом $p = 2m + 1$.

Пусть имеются дискретные во времени наблюдения над некоторым изучаемым процессом:

$$y_1, y_2, \ldots, y_i, \ldots, y_n,$$

где i – дискретный момент времени, равный порядковому номеру местоположения значения y_i в наборе данных;

n – объем выборки.

Тогда метод скользящей средней состоит в том, что исходный эмпирический временной ряд $y_1, \ldots, y_n$ преобразуется в ряд сглаженных значений (оценок) по формуле

$$\hat{y}_t = \frac{1}{p} \sum_{j=t-m}^{t+m} y_j,$$

где p – размер окна;
j – порядковый номер уровня в окне сглаживания;
m – величина, определяемая по формуле $m = (p-1)/2$.

Определение скользящей средней по четному числу членов ряда ($p = 2m$) несколько сложнее, поскольку вычисленное по аналогичной формуле усредненное значение нельзя сопоставить какому-либо определенному моменту времени t, так как средняя может быть отнесена только к середине между двумя датами, находящимися в середине окна сглаживания. Для определения сглаженных уровней при $p = 2m$ применяется так называемый *метод центрирования*, который заключается в нахождении средней из двух смежных скользящих средних для отнесения полученного уровня к определенной дате (*см.* пример 15.1).

При применении метода скользящей средней выбор размера окна сглаживания p должен осуществляться исходя из содержательных соображений и привязанности к периоду сезонности для сезонных волн. Если процедура скользящей средней используется для сглаживания несезонного ряда, то чаще всего размер окна сглаживания выбирают равным трем, пяти и семи. Чем больше размер окна, тем более гладкий вид имеет график скользящих средних.

Рассмотренный метод простой скользящей средней вполне приемлем, если графическое изображение временного ряда напоминает прямую линию. В этом случае не искажается динамика исследуемого явления. Однако когда тренд выравниваемого ряда имеет явно нелинейный характер и к тому же желательно сохранить мелкие волны, использовать для сглаживания ряда этот метод нецелесообразно, так как простая скользящая средняя может привести к значительным искажениям исследуемого процесса. В таких случаях более надежным является использование или метода взвешенной скользящей средней, или метода экспоненциального сглаживания.

Метод экспоненциального сглаживания*. Этот метод, как и метод скользящей средней, представляет собой некоторый способ усреднения значений эмпирического временного ряда $y_1, y_2, \ldots, y_i, \ldots, y_n$. В отличие от метода скользящей средней в определении экспоненциальной средней участвуют *все* наблюдения исходного временного ряда, но с разными весовыми коэффициентами (в методе простой скользящей средней все наблюдения временного ряда имеют вес, равный $1/p$). Экспоненциальная средняя обладает большей временной устойчивостью по сравнению со скользящей средней.

Для экспоненциального сглаживания момент времени, в который наблюдалось значение временного ряда, играет решающую роль. Здесь более старым наблюдениям приписываются экспоненциально убывающие веса, при этом в отличие от скользящего среднего учитываются *все* предшествующие наблюдения ряда, а не те, что попали в определенное окно. Формула метода простого экспоненциального сглаживания имеет следующий вид:

$$\hat{y}_t = (1-\alpha)\hat{y}_{t-1} + \alpha y_t,$$

где $0 < \alpha < 1$ – коэффициент экспоненциального сглаживания.

Когда эта формула применяется рекуррентно, то каждое новое теоретическое сглаженное значение вычисляется как взвешенное среднее текущего наблюдения и теоретического сглаженного значения предыдущего периода.

Очевидно, что результат сглаживания зависит от параметра α. Чем больше α, тем сильнее сказываются фактические наблюдаемые значения (при $\alpha = 1$ теоретические сглаженные значения предыдущего периода полностью игнорируются), чем меньше α, тем сильнее сказываются теоретические сглаженные значения (при $\alpha = 0$ полностью игнорируются фактические значения).

---

* Исторически метод экспоненциального сглаживания был независимо открыт Броуном и Холтом для решения задач прогнозирования спроса на запасные части вооружения и военной техники в интересах ВМС США.

15.2. Справочная информация по технологии работы

Режим работы «Скользящее среднее» служит для сглаживания уровней эмпирического временного ряда на основе метода простой скользящей средней.

Режим работы «Экспоненциальное сглаживание» служит для сглаживания уровней эмпирического временного ряда на основе метода простого экспоненциального сглаживания.

В диалоговых окнах данных режимов (рис. 15.1 и 15.2) задаются следующие параметры:

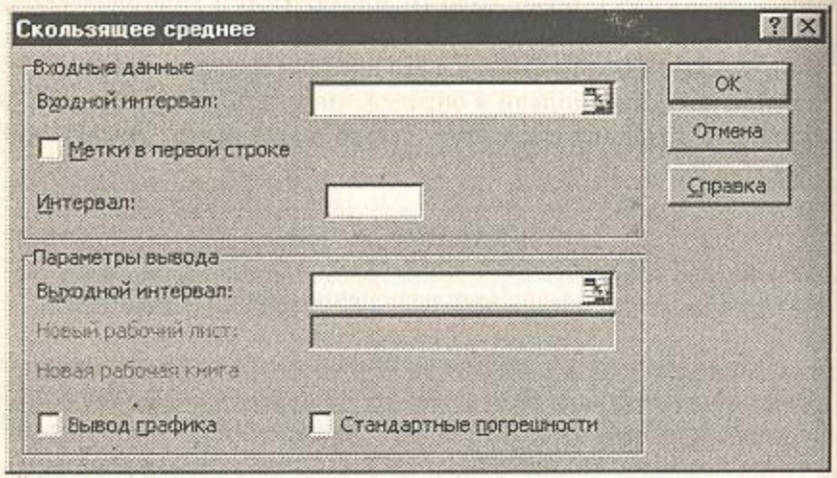

Рис. 15.1

1. *Входной интервал – см.* подразд. 1.1.2.
2. *Метки – см.* подразд. 1.1.2.
3. *Интервал* (только в диалоговом окне **Скользящее среднее**) – вводится размер окна сглаживания p. По умолчанию $p = 3$.
4. *Фактор затухания* (только в диалоговом окне **Экспоненциальное сглаживание**) – вводится значение коэффициента экспоненциального сглаживания α. По умолчанию $\alpha = 0{,}3$.
5. *Выходной интервал/Новый рабочий лист/Новая рабочая книга – см.* подразд. 1.1.2.

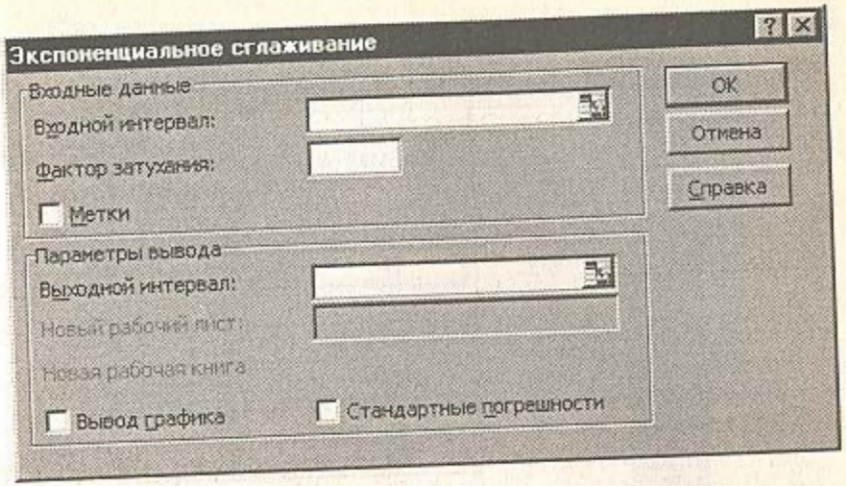

Рис. 15.2

6. *Вывод графика* – устанавливается в активное состояние для автоматической генерации на рабочем листе графиков фактических и теоретических уровней временного ряда.
7. *Стандартные погрешности* – устанавливается в активное состояние, если требуется включить в выходной диапазон столбец, содержащий стандартные погрешности.

Пример 15.1. Данные о среднедневной реализации (тыс. руб.) продуктов сельскохозяйственного производства магазинами потребительской кооперации города приведены в табл. 15.1, сформированной на рабочем листе Microsoft Excel [8].

В указанном периоде (1994–1997 гг.) требуется выявить основную тенденцию развития данного экономического процесса и характер его сезонных колебаний.

Для решения задачи используем режим работы «Скользящее среднее». Значения параметров, установленных в одноименном диалоговом окне, представлены на рис. 15.3, рассчитанные в данном режиме показатели – в табл. 15.2, а построенные графики – на рис. 15.4.

В столбце D (*см.* табл. 15.2) вычисляются значения сглаженных уровней. Например, значение первого сглаженного уровня рассчитывается в ячейке D6 по формуле =СРЗНАЧ(C3:C6), зна-

Таблица 15.1

	A	B	C
2	Год	Квартал	Размер реализации, тыс. руб.
3	1994	I	175
4		II	263
5		III	326
6		IV	297
7	1995	I	247
8		II	298
9		III	366
10		IV	341
11	1996	I	420
12		II	441
13		III	453
14		IV	399
15	1997	I	426
16		II	449
17		III	482
18		IV	460

чение второго сглаженного уровня – в ячейке D7 по формуле =СРЗНАЧ(C4:C7) и т. д.

В столбце E вычисляются значения стандартных погрешностей с помощью формулы =КОРЕНЬ(СУММКВРАЗН(блок __ фактических ___ значений; блок ___ прогнозных ___ значений)/размер ____ окна ____ сглаживания). Например, значение в ячейке E9 рассчитывается по формуле =КОРЕНЬ (СУММКВРАЗН(C6:C9;D6:D9)/4).

Вместе с тем, как отмечалось в подразд. 15.1, если размер окна сглаживания является четным числом ($p = 2m$), рассчитанное усредненное значение нельзя сопоставить какому-либо определенному моменту времени t, поэтому необходимо применять процедуру центрирования.

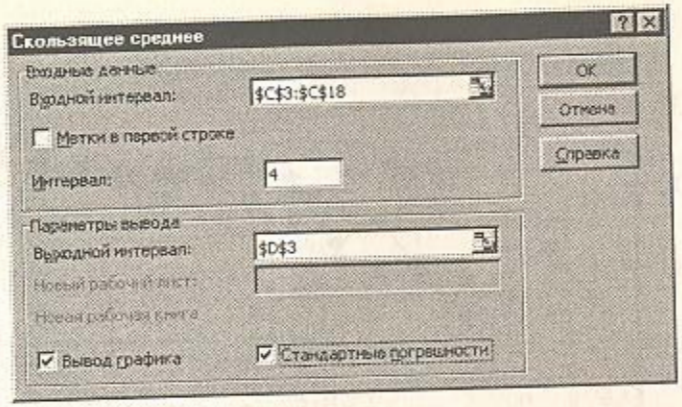

Рис. 15.3

Таблица 15.2

	A	B	C	D	E
2	Год	Квартал	Размер реализации, тыс. руб.	Сглаженные уровни	Стандартные погрешности
3	1994	I	175	#Н/Д	#Н/Д
4		II	263	#Н/Д	#Н/Д
5		III	326	#Н/Д	#Н/Д
6		IV	297	265,25	#Н/Д
7	1995	I	247	283,25	#Н/Д
8		II	298	292,00	#Н/Д
9		III	366	302,00	40,17
10		IV	297	313,00	39,47
11	1996	I	420	356,25	47,38
12		II	441	392,00	53,26
13		III	453	413,75	46,88
14		IV	399	428,25	47,07
15	1997	I	426	429,75	34,68
16		II	449	431,75	26,02
17		III	482	439,00	27,46
18		IV	460	454,25	23,42

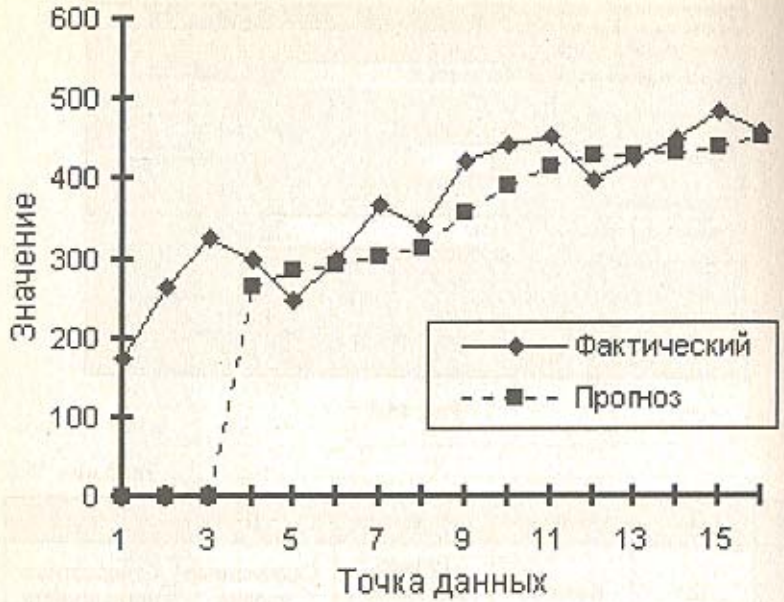

Рис. 15.4

Для рассматриваемого примера $p = 4$, поэтому процедура центрирования необходима. Так, первый сглаженный уровень (265,25) записывается между II и III кв. 1994 г., второй (283,25) – между III и IV кв. 1994 г. и т. д. Применяя процедуру центрирования (для этого используем функцию СРЗНАЧ), получаем *сглаженные уровни с центрированием*. Для III кв. 1994 г. определяется серединное значение между первым и вторым сглаженными уровнями: (265,25 + 283,25)/2 = 274,25; для IV кв. 1994 г. центрируются второй и третий сглаженные уровни: (283,25 + 292,00)/2 = 287,6 и т. д. Полученные значения новых сглаженных уровней представлены в табл. 15.3, а скорректированный график скользящей средней – на рис. 15.5.

Рассчитанные сглаженные уровни не только дают представление об общей тенденции поведения изучаемого временного ряда, но могут быть также использованы и для вычисления индексов сезонности I_S, совокупность которых характеризует сезонную волну

Таблица 15.3

	A	B	C	...	H	I
2	Год	Квартал	Размер реализации, тыс. руб.	...	Сглаженные уровни с центрированием	$\frac{y_t}{\hat{y}_t}$
3	1994	I	175	...		
4		II	263	...		
5		III	326	...	274,25	1,189
6		IV	297	...	287,63	1,033
7	1995	I	247	...	297,00	0,832
8		II	298	...	307,50	0,969
9		III	366	...	334,63	1,094
10		IV	341	...	374,13	0,911
11	1996	I	420	...	402,88	1,043
12		II	441	...	421,00	1,048
13		III	453	...	429,00	1,056
14		IV	399	...	430,75	0,926
15	1997	I	426	...	435,38	0,978
16		II	449	...	446,63	1,005
17		III	482	...		
18		IV	460	...		

исследуемого экономического процесса. Средние индексы сезонности определяются по формуле

$$\bar{I}_S = \frac{1}{u}\sum\frac{y_t}{\hat{y}_t},$$

где y_t – исходные уровни ряда;
$\hat{y}_t$ – сглаженные уровни ряда;
u – число одноименных периодов.

В табл. 15.3 (столбец I) представлены значения $y_t/\hat{y}_t$. Для получения средних индексов сезонности I_S производится осреднение исчисленных значений $y_t/\hat{y}_t$ по одноименным кварталам:

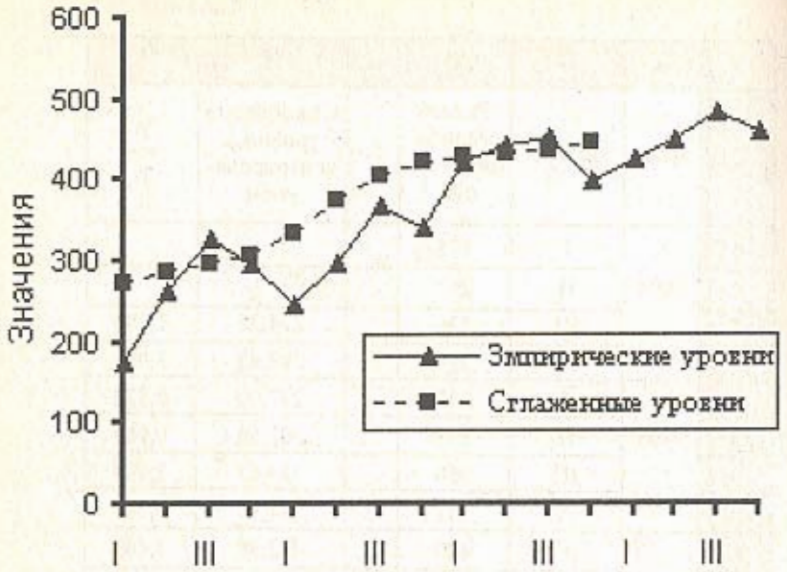

Рис. 15.5

I кв. – (0,832 + 1,043 + 0,978)/3 = 0,951, или 95,1%;
II кв. – (0,969 + 1,048 + 1,005)/3 = 1,007, или 100,7%;
III кв. – (1,189 + 1,094 + 1,056)/3 = 1,113, или 111,3%;
IV кв. – (1,033 + 0,911 + 0,926)/3 = 0,957, или 95,7%.

Исчисленные показатели являются средними индексами сезонных колебаний продажи сельскохозяйственной продукции по кварталам. Сезонная волна товарооборота сельскохозяйственной продукции (прирост в процентах к среднему уровню) изображена в виде столбиковой диаграммы на рис. 15.6.

Рассмотренная задача может быть решена и с помощью метода простого экспоненциального сглаживания. Для этого необходимо использовать режим работы «Экспоненциальное сглаживание». Значения параметров, установленных в одноименном диалоговом окне, представлены на рис. 15.7, рассчитанные в данном режиме показатели – в табл. 15.4, а построенные графики – на рис. 15.8.

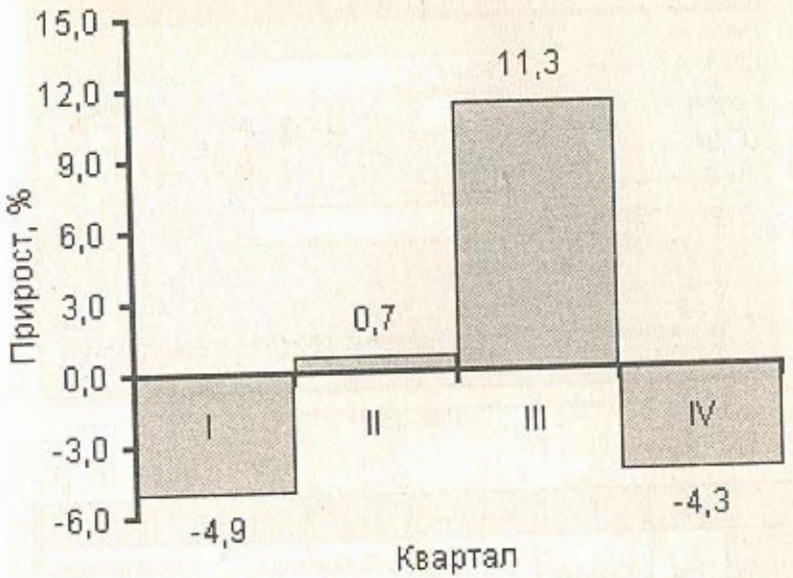

Рис. 15.6

В столбце F (*см.* табл. 15.4) вычисляются значения сглаженных уровней на основе рекуррентных соотношений. Например, значение первого сглаженного уровня рассчитывается в ячейке F4 по формуле = C3, значение второго сглаженного уровня – в ячейке F5 по формуле = 0,7 · C4+0,3 · F4, значение третьего сглаженного уровня – в ячейке F6 по формуле = 0,7 · C5+0,3 · F5 и т.д.

В столбце G рассчитываются значения стандартных погрешностей с помощью формулы =КОРЕНЬ(СУММКВРАЗН(блок __ фактических __ значений; блок __ прогнозных __значений)/3). Например, значение в ячейке G7 вычисляется по формуле = КОРЕНЬ(СУММКВРАЗН(C4:C6;F4:F6)/3).

Как легко заметить (сравните рис. 15.5 и 15.8), при использовании метода простого экспоненциального сглаживания в отличие от метода простой скользящей средней сохраняются мелкие волны.

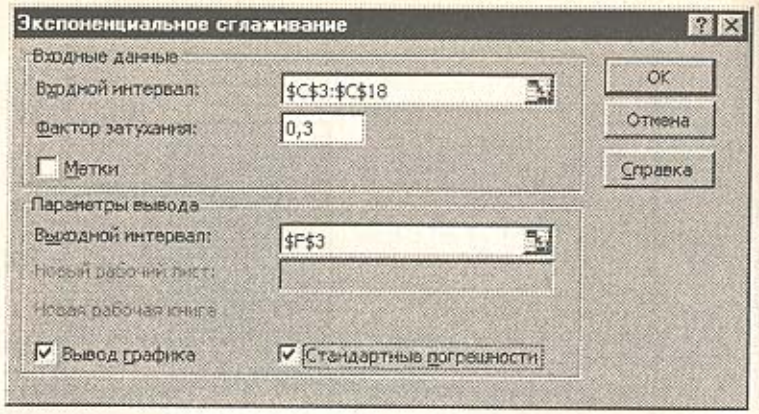

Рис. 15.7

Таблица 15.4

	A	B	C	...	F	G
2	Год	Квартал		...	Сглаженные уровни	Стандартные погрешности
3	1994	I		...	#Н/Д	#Н/Д
4		II		...	175,00	#Н/Д
5		III		...	236,60	#Н/Д
6		IV		...	299,18	#Н/Д
7	1995	I		...	297,65	72,44
8		II		...	262,20	59,34
9		III		...	287,26	35,84
10		IV		...	342,38	57,87
11	1996	I		...	341,41	49,95
12		II		...	396,42	64,23
13		III		...	427,63	52,17
14		IV		...	445,39	54,18
15	1997	I		...	412,92	39,93
16		II		...	422,07	31,45
17		III		...	440,92	31,87
18		IV		...	469,68	29,35

Рис. 15.8

ГЛАВА 16
Трендовые модели

16.1. Краткие сведения из теории статистики

Изложенные в главе 15 методы сглаживания временных рядов (метод скользящей средней и метод экспоненциального сглаживания) не дают теоретических рядов, в основе которых лежала бы определенная, математически выраженная закономерность изменения. Поэтому во многих случаях более результативным является применение *метода аналитического выравнивания*. Содержанием этого метода является то, что основная тенденция развития процесса (*тренд*) рассчитывается как функция времени

$$\hat{y}_t = f(t).$$

Теоретические уровни $\hat{y}_t$ определяются с использованием так называемой адекватной математической функции, которая наилучшим образом отображает основную тенденцию временного ряда. Подбор адекватной функции осуществляется методом наименьших квадратов (*см.* подразд. 14.1), при котором минимизируется сумма квадратов отклонений между эмпирическими y_t и теоретическими $\hat{y}_t$ уровнями ряда:

$$S = \sum_{i=1}^{n} (y_{t_i} - \hat{y}_{t_i})^2 \to \min.$$

Для оценки точности трендовой модели используют коэффициент детерминации

$$R^2 = \frac{\sigma_{\hat{y}}^2}{\sigma_y^2},$$

где $\sigma_{\hat{y}}^2 = \dfrac{\sum_{i=1}^{n} (\hat{y}_i - \bar{y})^2}{n}$ – дисперсия теоретических данных, полученных по трендовой модели;

$\sigma_y^2 = \dfrac{\sum_{i=1}^{n} (y_i - \bar{y})^2}{n}$ – дисперсия эмпирических данных.

Трендовая модель адекватна изучаемому процессу и отражает тенденцию его развития во времени при значениях R^2, близких к 1.

Важнейшей проблемой, требующей своего решения при применении метода аналитического выравнивания, является подбор математической функции, по которой рассчитываются теоретические уровни ряда. Если выбранный тип математической функции адекватен основной тенденции развития изучаемого процесса, то синтезированная трендовая модель может иметь полезное применение при изучении сезонных колебаний, прогнозировании и др.

Для обоснованного применения метода аналитического выравнивания в анализе временных рядов важно понимание сущности развития социально-экономических явлений во времени, знание их отличительных признаков.

В практике статистического изучения временных рядов различают следующие основные типы развития явлений во времени:

1) *равномерное развитие* – развитие с постоянным абсолютным приростом уровней временного ряда. Основная тенденция развития описывается линейным типом тренда:

$$\hat{y} = a_0 + a_1 t,$$

где a_0 – постоянная составляющая;

a_1 – коэффициент, характеризующий скорость (темп) развития изучаемого процесса и направление его развития (при $a_1 > 0$ уровни динамики равномерно возрастают, при $a_1 < 0$ – равномерно снижаются).

2) *равноускоренное (равнозамедленное) развитие* – развитие при постоянном увеличении (замедлении) темпа прироста уровней временного ряда. Основная тенденция развития описывается полиномом второй степени:

$$\hat{y} = a_0 + a_1 t + a_2 t^2,$$

где a_2 – коэффициент, характеризующий постоянное изменение скорости (темпа) развития (при $a_2 > 0$ происходит ускорение развития, при $a_2 < 0$ – замедление развития);

3) *развитие с переменным ускорением (замедлением)* – развитие при переменном увеличении (замедлении) темпа прироста уровней временного ряда. Основная тенденция описывается полиномом третьей степени:

$$\hat{y} = a_0 + a_1 t + a_2 t^2 + a_3 t^3,$$

где a_3 – коэффициент, характеризующий изменение ускорения развития (при $a_3 > 0$ ускорение возрастает, при $a_3 < 0$ – замедляется);

4) *развитие с замедлением роста в конце периода* – развитие, при котором прирост в конечных уровнях временного ряда стре-

21–2817

мится к нулю. Основная тенденция описывается логарифмической функцией

$$\hat{y} = a_0 + a_1 \ln t;$$

5) *развитие по экспоненте* – развитие, характеризующееся стабильным темпом роста (снижения). Основная тенденция описывается показательной (в частном случае экспоненциальной) функцией

$$\hat{y} = a_0 a_1^t,$$

где a_1 – коэффициент, характеризующий интенсивность развития.

6) *развитие по степенной функции* – развитие с постоянным относительным приростом уровней временного ряда. Основная тенденция развития описывается степенной функцией

$$\hat{y} = a_0 t^{a_1}.$$

Отметим, что пользоваться трендовыми моделями для краткосрочных и среднесрочных прогнозов следует только при выполнении следующих условий:

- период времени, за который изучается прогнозируемый процесс, должен быть достаточным для выявления закономерностей;
- трендовая модель в анализируемый период должна развиваться эволюционно;
- процесс, описываемый временным рядом, должен обладать определенной инерционностью, т. е. для наступления большого изменения в поведении процесса необходимо значительное время;
- автокорреляционная функция временного ряда и его остаточного ряда должна быть быстро затухающей, т. е. влияние более поздней информации должно сильнее отражаться на прогнозируемой оценке, чем влияние более ранней информации.

16.2. Справочная информация по технологии работы

В Microsoft Excel трендовые модели строятся на основе диаграмм, представляющих уровни динамики. Для эмпирического

временного ряда может быть построена диаграмма одного из следующих типов: гистограмма; линейчатая диаграмма; график; точечная диаграмма; диаграмма с областями.

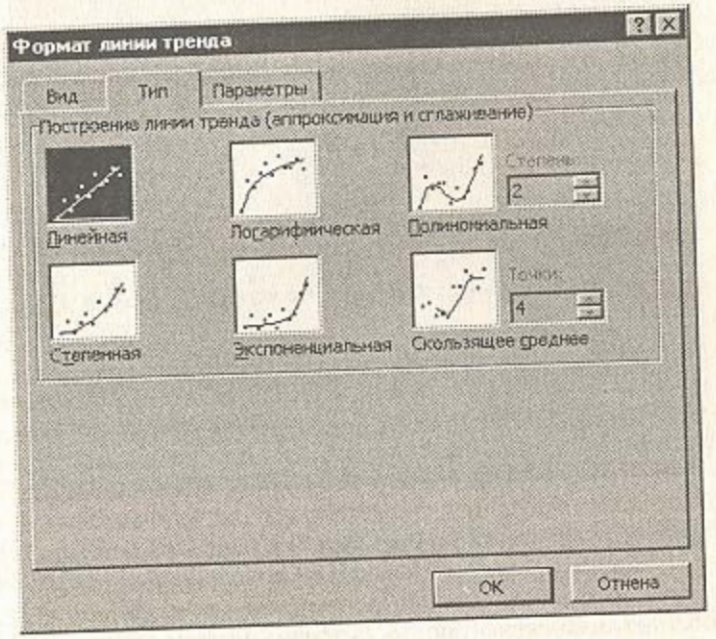

Рис. 16.1

Для построения линии тренда необходимо выделить временной ряд и выбрать в контекстном меню (вызывается щелчком правой клавиши мыши) команду **Добавить линию тренда**. Будет вызвано диалоговое окно **Линия тренда**, содержащее вкладку ***Тип*** (рис. 16.1), на которой задается тип тренда:

1) линейный;
2) логарифмический;
3) полиномиальный (от 2-й до 6-й степени включительно);
4) степенной;
5) экспоненциальный,
6) скользящее среднее (с указанием периода сглаживания от 2 до 15).

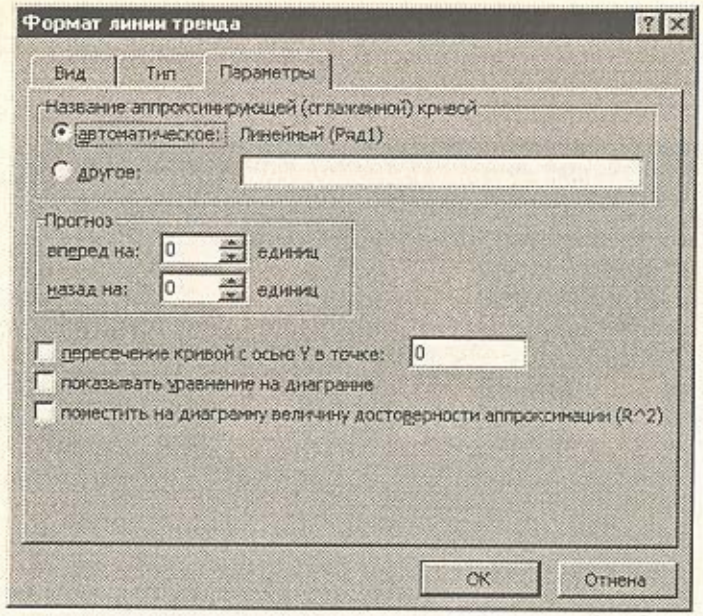

Рис. 16.2

Вкладка ***Параметры*** (рис. 16.2) предназначена для задания параметров тренда:

1. *Имя тренда* – имя линии тренда, располагается в легенде диаграммы; возможны следующие варианты задания имени тренда:

• *автоматическое* – Microsoft Excel именует линию тренда, основываясь на выбранном типе тренда и ряде динамики, с которым она ассоциирована, например, *Линейный (Ряд 1)*;

• *другое* – вводится уникальное имя тренда, максимальная длина составляет 256 символов.

2. *Прогноз вперед на* – количество периодов, на которое линия тренда проектируется в будущее, т. е. в направлении от оси *Y* (поле не доступно в режиме скользящего среднего).

3. *Прогноз назад на* – количество периодов, на которое линия тренда проектируется в прошлое, т. е. в направлении к оси *Y* (поле не доступно в режиме скользящего среднего).

4. *Пересечение кривой с осью Y в точке* – точка, в которой линия тренда пересекает ось *Y* (поле не доступно в режиме скользящего среднего).

5. *Показывать уравнение на диаграмме* – на диаграмме будет показано уравнение линии тренда.

6. *Поместить на диаграмму величину достоверности аппроксимации (R^2)* – на диаграмме будет показано значение коэффициента детерминации.

Наряду с линией тренда на графике временного ряда могут быть также изображены планки погрешностей.

Планки погрешностей используются во многих инженерных и статистических задачах для того, чтобы показать возможную погрешность значений эмпирического ряда (диапазон отклонений «плюс-минус» или в одну из сторон). В диаграммах планка погрешности изображается относительно значений эмпирического ряда.

Дополнить планками погрешностей ряды данных можно только для гистограмм, линейчатых диаграмм, графиков, диаграмм с областями и точечных диаграмм. *Y*-планки погрешностей отображаются вдоль оси значений *Y* (точечные диаграммы могут выводить также *X*-планки погрешностей вдоль оси *X*).

При изменении значений элементов ряда данных автоматически вычисляются новые величины погрешностей и соответствующим образом изменяются их планки.

Для вставки планок погрешностей следует выделить ряд данных и в контекстном меню выбрать команду **Формат рядов данных**. Будет вызвано диалоговое окно **Формат ряда данных**, содержащее вкладку ***Y-погрешности*** (рис. 16.3), которая обеспечивает выбор типа планок и варианта их расчета в зависимости от вида погрешности:

• *фиксированное значение* – за величину ошибки принимается заданное постоянное значение погрешностей;

• *относительное значение* – для каждой точки данных вычисляется отклонение на заданный процент;

• *стандартное отклонение* – вычисляется стандартное отклонение, которое затем умножается на заданное число (коэффициент кратности);

• *стандартная погрешность* – постоянная для всех элементов данных величина ошибки;

• *пользовательская* – вводится произвольный массив значений отклонений в положительную и/или отрицательную сторону (можно ввести ссылки на блок ячеек).

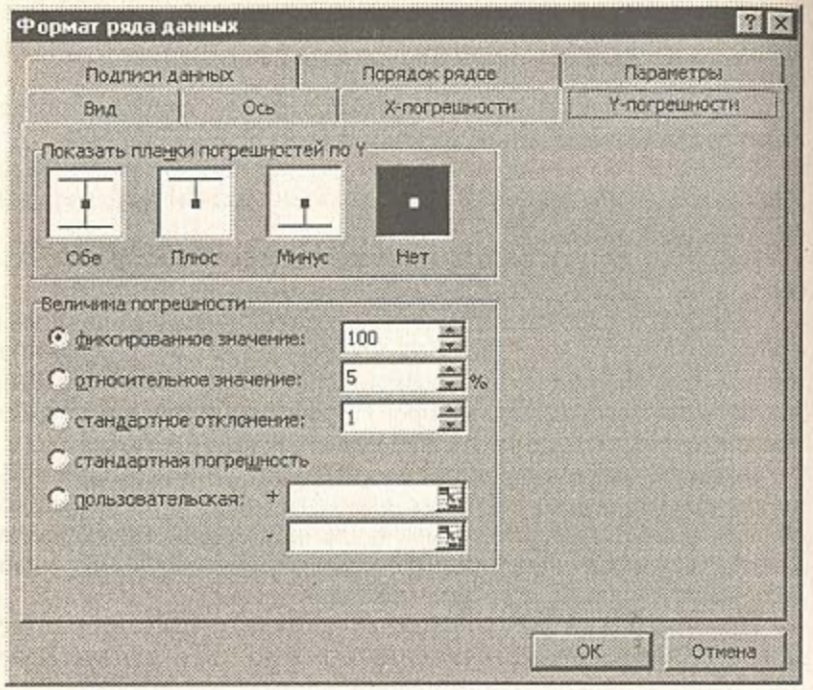

Рис. 16.3

Планки погрешности можно также форматировать. Для этого их следует выделить и выполнить команду контекстного меню **Формат полос погрешностей**.

Пример 16.1. Требуется по данным о розничном товарообороте региона (табл. 16.1) построить трендовую модель товарооборота [8].

Разнохарактерность изменений темпов роста (104,0 > 101,1 < < 107,7 > 102,7) и значительная колеблемость цепных абсолютных приростов (от 0,19 до 1,33) затрудняют определение типа динамики объема розничного товарооборота.

Для решения поставленной задачи, прежде всего в порядке первого приближения, намечаются типы функций, которые могут отобразить имеющиеся во временном ряду изменения. В помощь

Таблица 16.1

	B	C	D	E
3	Год	Объем розничного товарооборота, млрд руб.	Темп роста по годам, %	Абсолютный прирост по годам, млрд руб.
4	1985	16,4	–	–
5	1986	17,05	104,0	0,65
6	1987	17,24	101,1	0,19
7	1988	18,57	107,7	1,33
8	1989	19,08	102,7	0,51

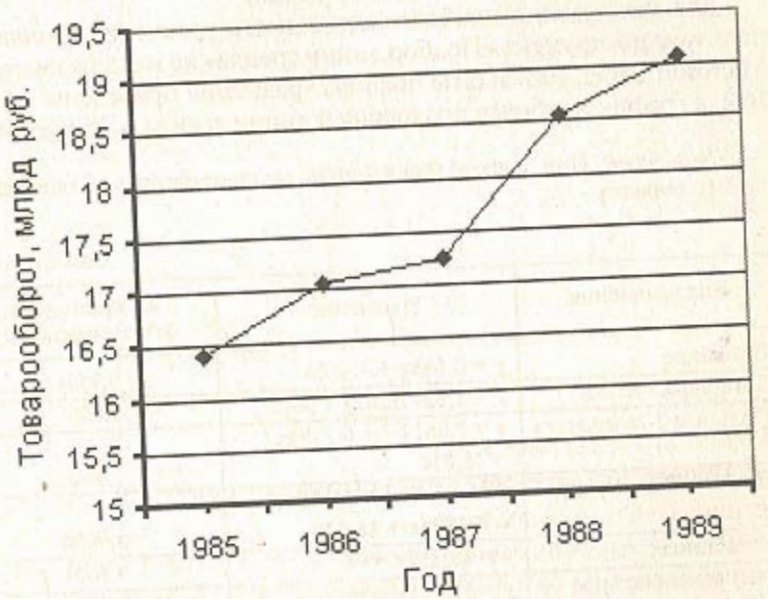

Рис. 16.4

этому исходные данные, приведенные в табл. 16.1, изображаются графически с помощью мастера диаграмм (рис. 16.4).

По характеру размещения уровней анализируемого временного ряда можно сделать предположение о возможном аналитическом выравнивании изучаемого ряда типовой математической функцией. Это может быть и линейная функция, и показательная, и полином 2-го порядка, и ряд других функций. Разнохарактерность темпов роста и значительная колеблемость цепных абсолютных приростов наталкивают на мысль, что развитие изучаемого процесса происходит с переменным ускорением, т. е. его основная тенденция описывается полиномом 3-го порядка:

$$\hat{y} = a_0 + a_1 t + a_2 t^2 + a_3 t^3.$$

Однако данная гипотеза требует количественного подтверждения, для чего необходимо осуществить перебор решений по намеченным типам математических функций.

Для нахождения наиболее адекватного уравнения тренда используем инструмент «Подбор линии тренда» из мастера диаграмм Microsoft Excel. Результаты подбора уравнения приведены в табл. 16.2, а график наиболее подходящей линии тренда – на рис. 16.5.

Примечание. При подборе уравнения не рассматривались полиномы выше 3-го порядка.

Таблица 16.2

Вид уравнения	Уравнение	Коэффициент детерминации R^2
Линейное	$y = 0{,}688x + 15{,}604$	0,9504
Логарифмическое	$y = 1{,}6245\mathrm{Ln}(x) + 16{,}113$	0,8561
Полином 2-го порядка	$y = 0{,}0614x^2 + 0{,}3194x + 16{,}034$	0,9610
Полином 3-го порядка	$\mathbf{y = -0{,}03x^3 + 0{,}3314x^2 - 0{,}3886x + 16{,}538}$	**0,9636**
Степенное	$y = 16{,}152x^{0{,}0921}$	0,8671
Экспоненциальное	$y = 15{,}701e^{0{,}0388x}$	0,9538

Принимая во внимание физическую сущность изучаемого процесса и результаты проведенного аналитического выравнивания (*см.* табл. 16.2), в качестве математической модели тренда выбираем полином 3-го порядка.

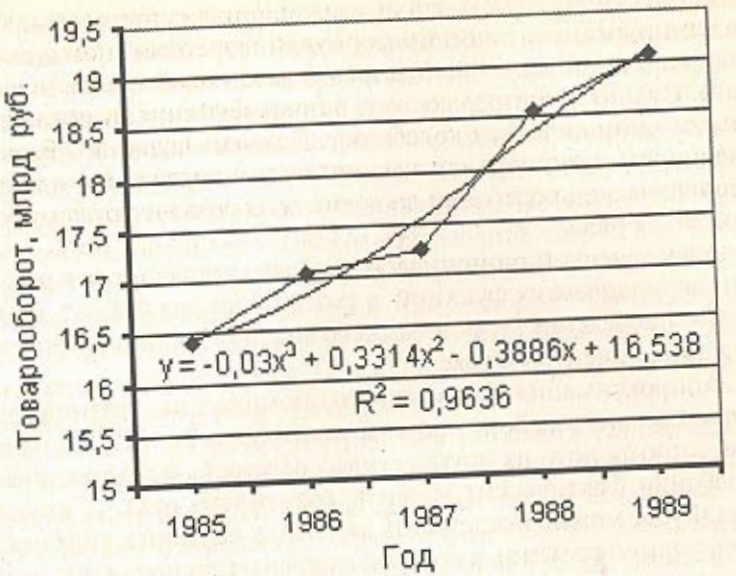

Рис. 16.5

ГЛАВА 17
Анализ Фурье

17.1. Краткие сведения из теории статистики

Как упоминалось в главе 16, при анализе экономических временных рядов наиболее часто в качестве трендовых моделей используются полиномы различных степеней, экспоненты, логистические кривые, кривые Гомперца и ряд других функций. Тем не менее моделирование временных рядов с помощью перечисленных функций не всегда дает удовлетворительные результаты, так как во временных рядах содержатся заметные периодические колебания вокруг общей тенденции или наблюдается автокоррелля-

ция не в самих уровнях, а в их отклонениях от полученных по определенным аналитическим формулам теоретических значений. В таких случаях следует использовать метод *гармонического анализа* ряда. Сущность метода состоит в представлении функций в виде суммы гармонических колебаний. Применительно к временным рядам целью данного анализа является выявление и измерение периодических колебаний во временных рядах и автокорреляции в остатках ряда.

Классический гармонический анализ заключается в разложении периодических функций в сходящийся ряд Фурье*. Практическое проведение гармонического анализа связано с вычислением коэффициентов Фурье.

Аппроксимация динамики экономических явлений рядом Фурье состоит в выборе таких гармонических колебаний, наложение которых друг на друга (сумма) отражало бы периодические колебания фактических уровней временного ряда. С помощью ряда Фурье можно представить динамику явлений в виде некоторой функции времени, в которой слагаемые расположены по убыванию периодов:

$$\hat{y}_t = a_0 + \sum(a_k \cos kt + b_k \sin kt).$$

В этом уравнении величина k определяет гармонику ряда Фурье и может быть взята целым числом (чаще всего от 1 до 4). Параметры уравнения определяются на основе метода наименьших квадратов и вычисляются по формулам:

$$a_0 = \frac{1}{N}\sum y_t; \qquad a_k = \frac{2}{N}\sum y \cos kt; \qquad b_k = \frac{2}{N}\sum y \sin kt.$$

*(Fourier Jean Baptiste Joseph) Фурье Жан Батист Жозеф (1768–1830) – французский математик и физик, иностранный почетный член Петербургской АН (1829), член Парижской академии наук (1817). Труды по алгебре, дифференциальным уравнениям и особенно по математической физике. Его «Аналитическая теория тепла» (1822) явилась отправным пунктом в создании теории тригонометрических рядов (рядов Фурье).

Исчисление параметров ряда Фурье может производиться и другими способами, в частности с помощью так называемого *преобразования Фурье*, которое применимо как к периодическим, так и непериодическим функциям.

Преобразование Фурье рассматривается в статистике обычно в рамках *одномерного спектрального анализа*, который является обобщенным случаем гармонического анализа. Теория спектрального анализа особенно широкое применение нашла в радиотехнических областях, где аппарат преобразования Фурье используется для преобразования сигналов или их корреляционных функций из временной области в частотную. Цель такого преобразования – решение задач фильтрации и прогнозирования с меньшим объемом математических вычислений.

При статистическом исследовании экономических процессов следует иметь в виду, что исходные данные имеют дискретный характер и могут быть представлены одним из двух вариантов:

- ограниченным дискретным набором данных, называемым в терминах спектрального анализа случайной последовательностью (реализацией);
- корреляционной функцией, описывающей дискретный экономический процесс.

Использование корреляционной функции возможно только при достаточно большом времени наблюдения, когда на основании существующей выборки данных обоснована стационарность этого процесса, т. е. неизменность во времени математического ожидания и дисперсии.

Из теории спектрального анализа для преобразования вышеназванных двух вариантов представления экономических процессов из временной в частотную область целесообразно заимствовать два понятия: спектр – для реализации случайной последовательности и спектральную плотность – для корреляционной функции случайного процесса.

Спектр – результат преобразования Фурье из временной области в частотную область конкретной реализации дискретного процесса (случайной последовательности).

Спектральная плотность – результат преобразования Фурье из временной области в частотную область корреляционной функции стационарного случайного процесса.

Рассмотрим некоторые понятия и определения спектрального анализа с целью его использования для исследования экономических процессов.

Спектральное разложение случайной функции $y(t)$ в действительной форме определяется выражением

$$y(t) = \sum_{k=0}^{\infty} (a_k \cos\omega_k t + b_k \sin\omega_k t),$$

где a_k, b_k – амплитуды для k-й гармоники;
ω_k – частота k-й гармоники.

Придадим спектральному разложению функции $y(t)$ в действительной форме комплексную форму. Комплексная форма записи удобна, в частности, потому, что всевозможные линейные операции над функциями, имеющими вид гармонических колебаний (дифференцирование, интегрирование, решение линейных дифференциальных уравнений и т. д.), осуществляются гораздо проще, когда эти гармонические колебания записаны не в виде синусов и косинусов, а в комплексной форме, в виде экспоненциальной функции. Для этого используем известные формулы Эйлера

$$\cos\omega_k t = \frac{e^{i\omega_k t} + e^{-i\omega_k t}}{2} \quad \text{и} \quad \sin\omega_k t = \frac{e^{i\omega_k t} - e^{-i\omega_k t}}{2i},$$

подставляя которые в формулу разложения функции $y(t)$ в действительной форме и осуществляя последующие преобразования, получаем итоговую формулу разложения функции $y(t)$ в комплексной форме:

$$y(t) = \sum_{k=-\infty}^{\infty} \Phi_k e^{i\omega_k t}.$$

При статистическом исследовании экономических процессов появляются достаточно серьезные ограничения, которые требуют

привлечения математического аппарата, несколько отличного от того, который был рассмотрен для случайной функции $y(t)$.

Во-первых, исходные данные дискретны, а значит, оперировать нужно не случайными функциями, а случайными последовательностями $y(n)$ ($n = t/N$, где T – период дискретизации случайной последовательности).

Во-вторых, набор исходных данных характеризуется ограниченным объемом, а это значит, что, используя терминологию спектрального анализа, следует оперировать случайными последовательностями $y(n)$ конечной длины N.

Для таких последовательностей вводится понятие *дискретного обратного преобразования Фурье* в виде суммы спектральных составляющих:

$$y(n) = \sum_{k=0}^{N-1} Y(k) e^{i\frac{2\pi}{N}kn},$$

где $Y(k)$ – комплексные числа из частотной области, соответствующие амплитудам k-й гармоники;
N – общее число наблюдений;
n – номер текущей точки.

Дискретное прямое преобразование Фурье позволяет вместо последовательности $y(n)$ из временной области получить комплексные числа $Y(k)$ в частотной области:

$$Y(k) = \sum_{n=0}^{N-1} y(n) e^{-i\frac{2\pi}{N}kn}.$$

Для уменьшения времени вычисления дискретного преобразования Фурье разработан алгоритм, получивший название *быстрого преобразования Фурье* (*БПФ*).

До середины 1960-х гг. для представления спектрального разложения использовались точные формулы, определяющие параметры синусов и косинусов. Соответствующие вычисления требовали, как минимум, N^2 комплексных умножений. Ситуация кардинально изменилась с открытием алгоритма БПФ, позволившего сделать время выполнения спектрального анализа ряда длины

N пропорциональным $N\log_2(N)$, что, конечно, является огромным прогрессом.

Вместе с тем следует отметить, что стандартный алгоритм БПФ обладает одним существенным недостатком: число данных ряда должно быть обязательно равным степени 2 (т. е. 16, 32, 64, 128, 256, ...). Один из путей преодоления этого недостатка – добавление в ряд констант (например, нулей) до тех пор, пока длина ряда не станет равной степени 2. Однако такой способ, применяемый при обработке электромагнитных сигналов, далеко не всегда приемлем для обработки данных, характеризующих экономические процессы. Поэтому перед применением алгоритма БПФ следует сформировать случайную последовательность $y(n)$ с длиной N, равной степени 2.

Примечание. В Microsoft Excel дискретное прямое преобразование Фурье реализовано без множителя $1/N$.

17.2. Справочная информация по технологии работы

Режим работы «Анализ Фурье» служит для реализации процедур дискретного прямого и дискретного обратного преобразований Фурье на основе стандартного алгоритма БПФ.

В диалоговом окне данного режима (рис.17.1) задаются следующие параметры:

1. *Входной интервал* – вводится ссылка на ячейки, содержащие данные, которые необходимо преобразовать. Входной диапазон может состоять из вещественных или комплексных данных (*см.* флажок *Инверсия*). Комплексные данные должны быть представлены в формате $x + y_j$ или $x + y_j$. Число значений во входном диапазоне должно быть равным степени 2. Максимальное число значений во входном диапазоне равно 4096.

2. *Метки – см.* подразд. 1.1.2.

3. *Выходной интервал/Новый рабочий лист/Новая рабочая книга – см.* подразд. 1.1.2.

4. *Инверсия* – данный флажок устанавливается в активное состояние для выполнения обратного преобразования Фурье и деактивизируется для выполнения прямого преобразования Фурье.

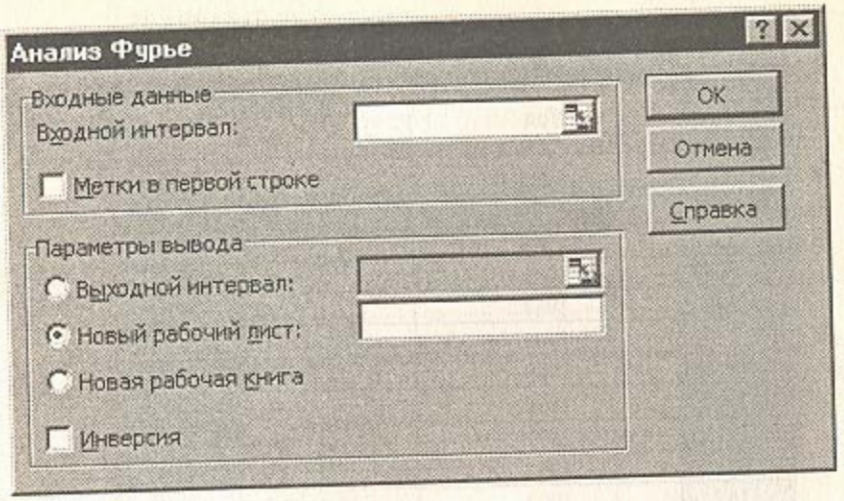

Рис. 17.1

Пример 17.1. Данные о динамике урожайности зерновых культур в одном из хозяйств области (ц/га) приведены в табл. 17.1, сформированной на рабочем листе Microsoft Excel.

Для представленного временного ряда требуется провести гармонический анализ динамики отклонений от основной тенденции.

Решение задачи начнем с построения трендовой модели ряда. В Microsoft Excel данную операцию удобнее всего проводить с помощью инструмента «Подбор линии тренда» из мастера диаграмм (порядок работы рассмотрен в подразд. 16.2).

Анализ трендовых моделей показывает, что в качестве рабочей модели можно выбрать линейную модель:

$$\hat{y}_t = -0{,}32t + 648{,}92.$$

Такой выбор обусловлен тем, что, во-первых, коэффициент детерминации $R^2 = 0{,}82$ имеет достаточно высокое значение (лишь очень незначительно уступает коэффициенту детерминации $R^2 = 0{,}85$ для полинома 2-го порядка); во-вторых, все коэффициенты модели значимы (*см.* табл. 17.4); в-третьих, при прочих равных условиях данная модель наиболее проста для вычислений и наиболее «прозрачна» для последующей экономической интерпретации.

Таблица 17.1

	B	C
2	Год	Урожайность, ц/га
3	1983	17,6
4	1984	18,1
5	1985	17,4
6	1986	16,8
7	1987	16,0
8	1988	15,4
9	1989	14,0
10	1990	16,6
11	1991	14,4
12	1992	14,2
13	1993	14,6
14	1994	13,8
15	1995	13,4
16	1996	14,2
17	1997	13,2
18	1998	13,2

График уравнения тренда показан на рис. 17.2.

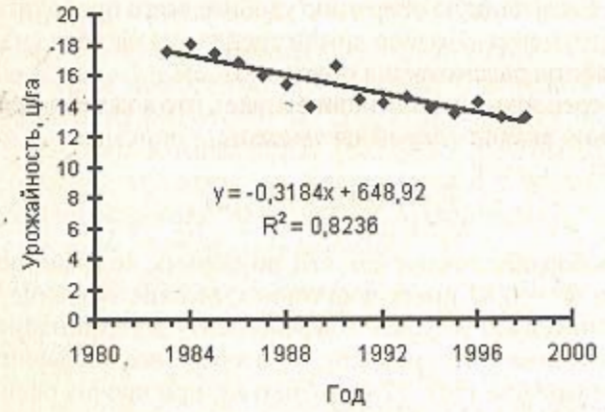

Рис. 17.2

Для более детального анализа построенной модели можно использовать режим «Регрессия» (*см.* главу 14). Показатели, рассчитанные в данном режиме, представлены в табл. 17.2–17.5.

Таблица 17.2

	B	C
21	ВЫВОД ИТОГОВ	
22		
23	*Регрессионная статистика*	
24	Множественный *R*	0,908
25	*R*-квадрат	0,824
26	Нормированный *R*-квадрат	0,811
27	Стандартная ошибка	0,726
28	Наблюдения	16

Таблица 17.3

	B	C	D	E	F	G
30	Дисперсионный анализ					
31		*df*	*SS*	*MS*	*F*	*Значимость F*
32	Регрессия	1	34,46	34,46	65,39	1,21E-06
33	Остаток	14	7,38	0,53		
34	Итого	15	41,84			

В табл. 17.5 (столбец *Остатки*) приведены значения отклонений от основной тенденции (разность между эмпирическими и теоретическими значениями). Гармонический анализ вычисленных отклонений проведем с помощью режима «Анализ Фурье». Значения параметров, установленных в одноименном диалоговом окне, представлены на рис. 17.3, а рассчитанные в данном режиме показатели – в табл. 17.6 (столбец E).

22–2817

Таблица 17.4

	B	C	D	E	F	G	H	I	J
36		*Коэффициенты*	*Стандартная ошибка*	*t-статистика*	*P-значение*	*Нижние 95%*	*Верхние 95%*	*Нижние 95,0%*	*Верхние 95,0%*
37	Y-пересечение	648,92	78,37	8,28	9,16E-07	480,83	817,02	480,83	817,02
38	Переменная	–0,32	0,04	–8,09	1,21E-06	–0,40	–0,23	–0,40	–0,23

Таблица 17.5

	B	C	D
44	*Наблюдение*	*Предсказанное*	Остатки
45	1	17,57	0,03
46	2	17,25	0,85
47	3	16,93	0,47
48	4	16,61	0,19
49	5	16,30	– 0,30
50	6	15,98	– 0,58
51	7	15,66	– 1,66
52	8	15,34	1,26
53	9	15,02	– 0,62
54	10	14,70	– 0,50
55	11	14,39	0,21
56	12	14,07	– 0,27
57	13	13,75	– 0,35
58	14	13,43	0,77
59	15	13,11	0,09
60	16	12,79	0,41

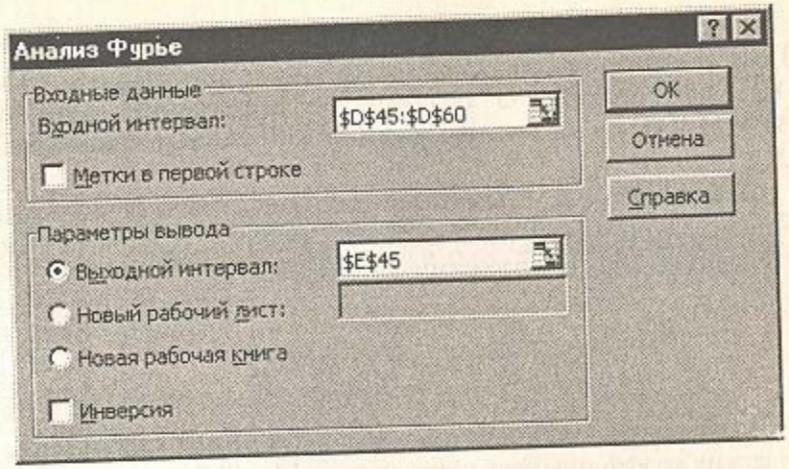

Рис. 17.3

Таблица 17.6

	D	E	F	G
44	*Остатки*	*Комплексные числа*	*Действительная часть Y_d*	*Мнимая часть Y_m*
45	0,03	0	0,000	
46	0,85	3,21792468331082 + 0,985461907442038i	3,218	0,985
47	0,47	1,39644406067929 – 1,12572977811612i	1,396	–1,126
48	0,19	–2,23291382828564 – 1,27069167046029i	–2,233	–1,271
49	–0,30	–0,347058823574952 + 1,04705882357496i	–0,347	1,047
50	–0,58	0,710369056389691 – 3,48957546369093i	0,710	–3,490
51	–1,66	–1,29056170782943 + 3,38015257473397i	–1,291	3,380
52	1,26	0,916384794285294 – 1,02165718008843i	0,916	–1,022
53	–0,62	–4,24705882357485	–4,247	0,000
54	–0,50	0,916384794285297 + 1,02165718008842i	0,916	1,022
55	0,21	–1,29056170782943 – 3,38015257473397i	–1,291	–3,380
56	–0,27	0,710369056389701 + 3,48957546369093i	0,710	3,490
57	–0,35	–0,347058823574954 – 1,04705882357496i	–0,347	–1,047
58	0,77	–2,23291382828563 + 1,2706916704603i	–2,233	1,271
59	0,09	1,39644406067929 + 1,12572977811611i	1,396	1,126
60	0,41	3,21792468331082 – 0,985461907442042i	3,218	–0,985

В столбце F с помощью инженерной функции МНИМ.ВЕЩ рассчитаны действительные части комплексных чисел (Y_d), а в столбце G с помощью инженерной функции МНИМ.ЧАСТЬ вычислены мнимые части комплексных чисел (Y_m).

Действительные и мнимые части рассчитанных в режиме «Анализ Фурье» комплексных чисел связаны с гармоническими коэффициентами следующими соотношениями:

$$a_0 = \frac{Y_{d_0}}{N},\quad a_k = \frac{Y_{d_k}}{N},\quad b_k = \frac{Y_{m_k}}{N}.$$

Значения рассчитанных по указанным соотношениям гармонических коэффициентов приведены в табл. 17.7.

Таблица 17.7

	H	I	J	K
45	a_0	0,00		
46	a_1	0,40	b_1	–0,12
47	a_2	0,09	b_2	0,14
48	a_3	–0,14	b_3	0,16
49	a_4	–0,02	b_4	–0,13
50	a_5	0,04	b_5	0,44
51	a_6	–0,08	b_6	–0,42
52	a_7	0,06	b_7	0,13
53	a_8	–0,27	b_8	0,00
54	a_9	0,06	b_9	–0,13
55	a_{10}	–0,08	b_{10}	0,42
56	a_{11}	0,04	b_{11}	–0,44
57	a_{12}	–0,02	b_{12}	0,13
58	a_{13}	–0,14	b_{13}	–0,16
59	a_{14}	0,09	b_{14}	–0,14
60	a_{15}	0,20	b_{15}	0,12

Для нахождения теоретических значений $\hat{y}_t$ необходимо от реального времени перейти к «радианному» времени по формуле

$$t_n = \frac{2\pi n}{N}.$$

В табл. 17.8 приведены значения «радианного» времени (столбец F), теоретические значения первых четырех гармоник (столбцы GJ) и их итоговая сумма (столбец K), соответствующая гармонической модели:

$$\hat{y}_t = a_0 + a_1\cos t + b_1\sin t + a_2\cos 2t + b_2\sin 2t + a_3\cos 3t + b_3\sin 3t + a_4\cos 4t + b_4\sin 4t.$$

Таблица 17.8

	E	F	G	H	I	J	K
63	n	t	U_1	U_2	U_3	U_4	$\hat{y}_t$
64	0	0,000	0,402	0,087	–0,140	–0,022	0,328
65	1	0,393	0,324	0,161	0,093	–0,131	0,448
66	2	0,785	0,197	0,141	0,211	0,022	0,571
67	3	1,178	0,040	0,038	0,068	0,131	0,277
68	4	1,571	–0,123	–0,087	–0,159	–0,022	–0,391
69	5	1,963	–0,268	–0,161	–0,190	–0,131	–0,750
70	6	2,356	–0,372	–0,141	0,014	0,022	–0,477
71	7	2,749	–0,419	–0,038	0,200	0,131	–0,126
72	8	3,142	–0,402	0,087	0,140	–0,022	–0,197
73	9	3,534	–0,324	0,161	–0,093	–0,131	–0,387
74	10	3,927	–0,197	0,141	–0,211	0,022	–0,246
75	11	4,320	–0,040	0,038	–0,068	0,131	0,060
76	12	4,712	0,123	–0,087	0,159	–0,022	0,173
77	13	5,105	0,268	–0,161	0,190	–0,131	0,165
78	14	5,498	0,372	–0,141	–0,014	0,022	0,239
79	15	5,890	0,419	–0,038	–0,200	0,131	0,312

На рис. 17.4 представлены эмпирический график отклонений урожайности от основной тенденции (ряд 1) и теоретические графики первых четырех гармоник (ряды 2–5). На рис. 17.5 этот же эмпирический график показан вместе с итоговым теоретическим графиком (ряд 6), полученным в результате суммирования первых четырех гармоник (рядов 2–5).

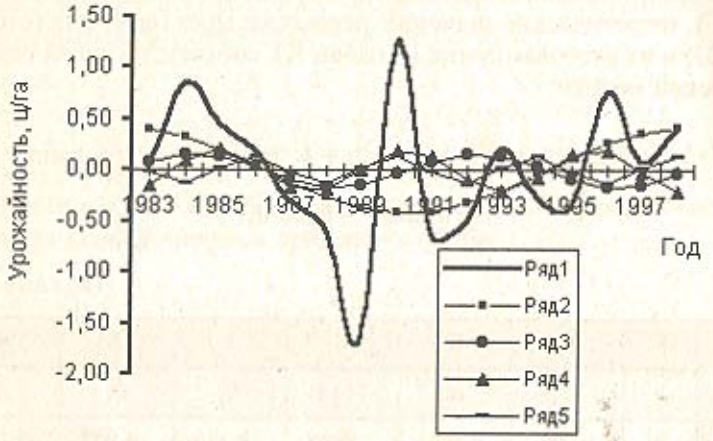

Рис. 17.4

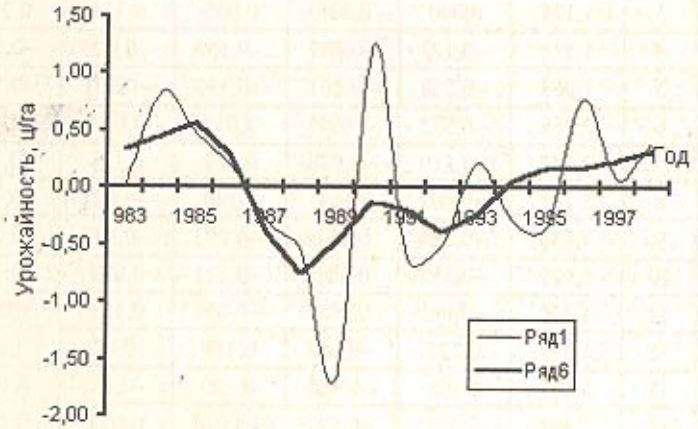

Рис. 17.5

ПРИЛОЖЕНИЕ

Совместное использование режимов надстройки «Пакет анализа»

При исследовании многих социально-экономических явлений и процессов часто приходится использовать не один, а несколько методов статистического анализа данных, что позволяет наиболее полно раскрыть их сущность, закономерности и тенденции развития. При этом необходимый набор и порядок применения статистических методов определяется исходя из цели исследования и характера решаемых задач.

Технологию совместного применения нескольких режимов работы надстройки «Пакет анализа» в ходе проведения комплексного статистического исследования рассмотрим на одном из примеров, встречающихся в приборостроительной практике.

Пример П1. Предприятие «Импульс» за месяц произвело 2000 приборов, которым были присвоены заводские номера с 5001 по 7000 включительно. Все приборы изготавливаются по технической документации, в соответствии с которой дисперсия чувствительности приборов не превышает 25 мкВ^2/м^2.

Требуется на основе выборочного обследования сделать заключение о характеристиках приборов всей партии, установить степень и характер зависимости предельной частоты распознаваемого прибором сигнала и его чувствительности, выяснить, влияет ли на значение чувствительности тип встраиваемой в прибор ферритовой антенны.

Для ответа на поставленные вопросы в первую очередь должна быть сформирована выборочная совокупность, обладающая свойством репрезентативности, что, в свою очередь, требует первоначального определения необходимого объема выборки (*см.* главу 3).

Необходимый объем выборки рассчитывается по формуле

$$n = \frac{t^2\sigma^2}{\Delta_x^2},$$

где Δ_x – предельная ошибка выборки;
σ^2 – дисперсия генеральной совокупности;
t – коэффициент доверия (определяется в зависимости от того, с какой доверительной вероятностью нужно гарантировать результаты выборочного обследования).

Для формирования контрольной выборки используем схему случайного повторного отбора при условии, что предельная ошибка выборки не превышает 3 мкВ/м с уровнем надежности не менее 95 %.

Подставляя исходные данные задачи, рассчитываем необходимый объем контрольной выборки:

$$n = \frac{t^2\sigma^2}{\Delta_x^2} = \frac{1{,}96^2 \cdot 25}{3^2} = 10{,}68 \approx 11 \quad \text{(приборов)}.$$

Примечание. В расчете необходимого объема выборки используется коэффициент доверия t, для вычисления которого в Microsoft Excel предусмотрена функция СТЬЮДРАСПОБР (*см.* подразд. 6.3.8). Здесь коэффициент доверия t определяется по формуле = СТЬЮДРАСПОБР (0,05; 1999), где 0,05 =1–0,95 – требуемый уровень значимости, 1999 = 2000-1 – число степеней свободы.

Таким образом, минимально допустимый объем выборки составляет 11 приборов. При меньшем объеме выборка не будет репрезентативной.

После определения минимально допустимого объема выборки на рабочем листе Microsoft Excel в диапазон размером в 2000 ячеек (например, в B1:B2000) введем заводские номера приборов с 5001 по 7000 и сформируем контрольную выборку (*см.* главу 3).

Для быстрого ввода исходных данных (объем генеральной совокупности составляет все же 2000 ед.!) рекомендуем использовать такой технический прием, как копирование ячеек с помощью правой клавиши мыши с последующей установкой через контекстное меню *арифметической прогрессии с шагом 1.*

В результате будет сформирована выборка из 11 приборов с заводскими номерами: 5141, 5155, 5349, 5460, 5565, 5706, 5714, 5768, 6501, 6771, 6972.

Отобранные приборы прошли стендовые испытания, на которых были определены тактико-технические характеристики каждого прибора (табл. П.1).

Таблица П.1

	D	E	F	G
4	Заводской номер прибора	Тип ферритовой антенны	Чувствительность прибора, мкВ/м	Частота распознаваемого сигнала, МГц
5	5141	ФА-77	90	10,07
6	5155	ФА-77м	96	9,73
7	5349	ФА-77	92	10,04
8	5460	ФА-77м	98	9,82
9	5565	ФА-77	86	10,57
10	5706	ФА-77м	88	10,02
11	5714	ФА-77м	98	9,67
12	5768	ФА-77	90	9,98
13	6501	ФА-77	86	10,51
14	6771	ФА-77	92	9,92
15	6972	ФА-77м	90	9,93

На основании полученных из контрольной выборки значений характеристик приборов сделаем заключение о чувствительности приборов всей партии. Для этого с помощью режима «Описательная статистика» (*см.* главу 4) рассчитаем показатели, представленные в табл. П.2.

Во-первых, убедимся, что дисперсия чувствительности приборов не превышает 25 мкВ^2/м^2 (показатель *Дисперсия выборки*), а предельная ошибка выборки – 3 мкВ/м (показатель *Уровень надежности*).

Во-вторых, на основании рассчитанных по контрольной выборке показателей (*см.* табл. П. 2) с уровнем надежности 95 % можно предположить, что средняя чувствительность приборов всей партии будет находиться в пределах от 88,56 (=91,45–2,89) мкВ/м до 94,34 (=91,45+2,89) мкВ/м (пояснения к расчетам *см.* в подразд. 4.2).

Таблица П. 2

	H	I
4	*Чувствительность*	
5		
6	Среднее	91,45
7	Стандартная ошибка	1,30
8	Медиана	90,00
9	Мода	90,00
10	Стандартное отклонение	4,30
11	Дисперсия выборки	18,47
12	Эксцесс	–0,93
13	Асимметричность	0,43
14	Интервал	12,00
15	Минимум	86,00
16	Максимум	98,00
17	Сумма	1006,00
18	Счет	11,00
19	Наибольший(1)	98,00
20	Наименьший(1)	86,00
21	Уровень надежности(95,0%)	2,89

В-третьих, коэффициент вариации

$$v = \frac{\sigma}{\bar{x}} \cdot 100\,\% = \frac{4,30}{91,45} \cdot 100\,\% \approx 4,7\%$$

существенно меньше 40%, что свидетельствует о малой колеблемости признака в исследованной выборочной совокупности. Надежность средней подтверждается также и ее незначительным отклонением от медианы: 91,45 – 90,00 = 1,45.

В-четвертых, незначительное положительное значение коэффициента асимметрии A_s позволяет говорить о том, что данное эмпирическое распределение имеет несущественную правостороннюю асимметрию, а отрицательное значение эксцесса E_k – о его плосковершинности, т.е. об отсутствии скопления членов ряда в центре распределения.

Следующим этапом проводимого исследования является установление степени и характера взаимосвязи предельной частоты распознаваемого прибором сигнала Y и его чувствительности X.

Для оценки тесноты связи между двумя величинами чаще всего используются коэффициенты ковариации и корреляции (*см.* главу 13). Результаты расчетов этих коэффициентов приведены соответственно в табл. П.3 и П.4.

Таблица П.3

	E	F	G
68		Чувствительность	Частота
69	Чувствительность	18,47	
70	Частота	–1,05	0,08

Таблица П. 4

	E	F	G
73		Чувствительность	Частота
74	Чувствительность	1	
75	Частота	–0,86	1

Как видим, связь между предельной частотой распознаваемого прибором сигнала Y и его чувствительностью X является высокой и обратной ($r_{xy} = -0,86$), т. е. с повышением чувствительности прибора предельная частота распознаваемого им сигнала уменьшается.

Кроме того, для установления степени взаимосвязи между двумя величинами можно также использовать ранговый коэффициент

Спирмена. Для расчета этого коэффициента используем режим работы «Ранг и персентиль» (*см.* главу 5), результаты выполнения которого представлены в табл. П.5.

Таблица П. 5

	D	E	F	G	H	I	J	K
35	*Точка*	*Столбец1*	*Ранг*	*Процент*	*Точка*	*Столбец2*	*Ранг*	*Процент*
36	4	98	1	90,00%	5	10,57	1	100,00%
37	7	98	1	90,00%	9	10,51	2	90,00%
38	2	96	3	80,00%	1	10,07	3	80,00%
39	3	92	4	60,00%	3	10,04	4	70,00%
40	10	92	4	60,00%	6	10,02	5	60,00%
41	1	90	6	30,00%	8	9,98	6	50,00%
42	8	90	6	30,00%	11	9,93	7	40,00%
43	11	90	6	30,00%	10	9,92	8	30,00%
44	6	88	9	20,00%	4	9,82	9	20,00%
45	5	86	10	,00%	2	9,73	10	10,00%
46	9	86	10	,00%	7	9,67	11	,00%

По данным сгенерированной табл. П. 5 заполняем графы *Ранг* R_X и *Ранг* R_Y (табл. П. 6), на основании которых производим вычисления квадратов разности рангов (d_i^2).

На заключительном этапе вычисляем коэффициент Спирмена по формуле

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)},$$

подставляя в которую исходные и рассчитанные данные, получим

$$\rho = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)} = 1 - \frac{6 \cdot 385}{11 \cdot (11^2-1)} = -0{,}75.$$

Таблица П.6

	B	C	D	E	F	G
50	Номер прибора	Чувствительность X, мкВ/м	Частота Y, МГц	Ранг R_X	Ранг R_Y	Квадрат разности рангов $d_i^2 = (R_X - R_Y)^2$
51	5141	90	10,07	6	3	9
52	5155	96	9,73	3	10	49
53	5349	92	10,04	4	4	0
54	5460	98	9,82	1	9	64
55	5565	86	10,57	10	1	81
56	5706	88	10,02	6	5	1
57	5714	98	9,67	1	11	100
58	5768	90	9,98	6	6	0
59	6501	86	10,51	10	2	64
60	6771	92	9,92	4	8	16
61	6972	90	9,93	6	7	1
62					Σ =	385

Пользуясь шкалой Чеддока (*см.* подразд. 13.1), можно констатировать, что теснота связи между чувствительностью прибора X и значением предельной распознаваемой им частоты Y является высокой, что подтверждает сделанный ранее вывод.

Значительно более сложной задачей является определение аналитического выражения связи между величинами X и Y, т. е. нахождение вида уравнения регрессии, наиболее подходящего для описания исследуемого явления. Здесь в первую очередь следует принимать во внимание физическую суть явления. Если исследователь такой информацией не располагает, то единственным подходом остается последовательный перебор основных видов уравнений (линейное, логарифмическое, экспоненциальное, полином 2-го порядка и т. п.).

Допустим, что в рассматриваемой ситуации не известен предполагаемый вид уравнения зависимости предельной частоты Y от чувствительности прибора X. Для нахождения такого уравнения используем инструмент «Подбор линии тренда» из мастера диаграмм (*см.* под-

разд. 16.2). Результаты подбора уравнения приведены в табл. П. 7, а график наиболее подходящего уравнения – на рис. П.1.

Таблица П. 7

Вид уравнения	Уравнение	Коэффициент детерминации R^2
Линейное	$y = 0{,}06x + 15{,}23$	0,74
Логарифмическое	$y = -5{,}28\ln(x) + 33{,}88$	0,75
Полином 2-го порядка	$y = 0{,}007x^2 - 1{,}28x + 71{,}57$	0,88
Полином 3-го порядка	$\mathbf{y = -0{,}001x^3 + 0{,}38x^2 - 35{,}07x + 1102{,}1}$	**0,91**
Степенное	$y = 106{,}55x^{-0{,}52}$	0,76
Экспоненциальное	$y = e^{-0{,}006x}$	0,75

Примечание. При подборе уравнения не рассматривались полиномы выше 3-го порядка.

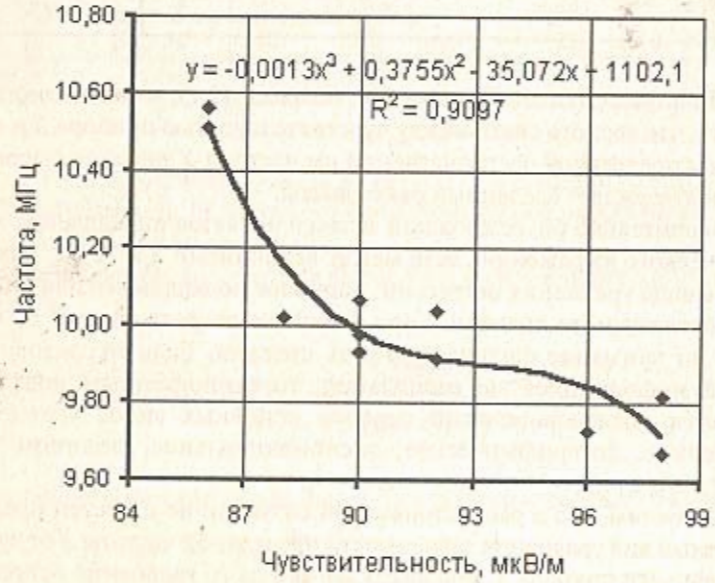

Рис. П.1

Последним этапом проводимого исследования является подтверждение (или опровержение) предположения о влиянии типа встраиваемой в прибор ферритовой антенны на чувствительность прибора.

При сборке приборов используются два типа антенн: ФА-77 и ФА-77м. Типы антенн в приборах контрольной выборки приведены в табл. П.1.

Для подтверждения (или опровержения) предположения о влиянии типа антенны на чувствительность прибора используем режим работы «Однофакторный дисперсионный анализ» (*см.* главу 11). Для этого исходные данные сгруппируем по типам антенн (табл. П.8), после чего произведем вычисления. Рассчитанные показатели представлены в табл. П.9 и П.10 (при уровне значимости α=0,05).

Таблица П.8

	E	F
82	Чувствительность приборов контрольной выборки по типам антенн	
83	ФА-77	ФА-77м
84	90	96
85	92	98
86	86	88
87	90	98
88	86	90
89	92	

Таблица П.9

	D	E	F	G	H
97	Однофакторный дисперсионный анализ				
98					
99	ИТОГИ				
100	*Группы*	*Счет*	*Сумма*	*Среднее*	*Дисперсия*
101	ФА-77	6	536	89,33	7,47
102	ФА-77м	5	470	94,00	22,00

Таблица П. 10

	B	C	D	E	F	G	H
105	Дисперсионный анализ						
106	*Источник вариации*	*SS*	*df*	*MS*	*F*	*P-значение*	*F критическое*
107	Между группами	59,39	1	59,39	4,26	0,07	5,12
108	Внутри групп	125,33	9	13,93			
109							
110	Итого	184,73	10				

Из табл. П.10 находим, что расчетное значения F-критерия F_p = = 4,26, а критическая область образуется правосторонним интервалом (5,12; +∞). Так как F_p не попадает в критическую область, то предположение о влиянии типа антенны на чувствительность прибора отвергаем.

Следующий пример демонстрирует возможность использования мастера диаграмм для решения простых оптимизационных задач с одной переменной.

Пример П.2. Небольшое частное кондитерское предприятие занимается производством фирменного печенья. Постоянные издержки производства (*FC*) составляют 20000 руб. в месяц, а средние переменные издержки (*AVC*) – 12 руб. на один килограмм произведенной продукции.

Если предприятие будет производить Q кг печенья в месяц, то его общие издержки (*TC*) составят 20000+12Q руб., а общая выручка (*TR*) – QP руб., где P – цена одного килограмма печенья (все произведенное печенье продается). Тогда прибыль предприятия (*Pr*) составит TR-TC = QP-20000-12Q руб.

С целью получения максимальной прибыли предприятие каждый месяц изменяет цену P и анализирует, как на изменение цены реагирует спрос населения. Сведения о цене и объеме продаж за первые шесть месяцев деятельности предприятия, а также рассчитанные на их основе показатели издержки, выручки и прибыли представлены в табл. П. 11, сформированной на рабочем листе Microsoft Excel.

Таблица П.11

	B	C	D	E	F	G	H
4		январь	февраль	март	апрель	май	июнь
5	Цена, руб./кг	20,00	20,50	23,50	23,00	22,50	21,80
6	Продано, кг	6300	6400	4850	4950,00	5500,00	5700,00
7	Общая выручка, руб.	126000,00	131200,00	113975,00	113850,00	123750,00	124260,00
8	Общие издержки, руб.	95600,00	96800,00	78200,00	79400,00	86000,00	88400,00
9	Прибыль, руб.	30400,00	34400,00	35775,00	34450,00	37750,00	35860,00

По имеющимся данным требуется определить, по какой цене и в каком объеме следует производить продукцию в июле, чтобы получить наибольшую прибыль.

Идея решения задачи состоит в нахождении аналитической зависимости между спросом населения и ценой на покупаемую продукцию, т. е. в нахождении вида функции *Спрос=f(Цена)*.

Решим рассматриваемую задачу с помощью мастера диаграмм.

На основе диапазона данных C5:H6 построим точечную диаграмму и, используя средства форматирования, приведем ее к удобному для восприятия виду. На построенной диаграмме выделим ряд значений и, вызвав контекстное меню (вызывается при нажатии правой клавиши мыши), выберем команду **Добавить линию тренда**. Будет вызвано диалоговое окно **Линия тренда**, содержащее вкладку ***Тип*** (*см.* рис. 16.1), где задается вид тренда (уравнения): линейный; логарифмический; полиномиальный (от 2-й до 6-й степени включительно); степенной; экспоненциальный, скользящее среднее (с указанием периода сглаживания от 2 до 15).

Здесь сразу же следует заметить, что при выборе вида приближающего уравнения, прежде всего, должна приниматься во внимание экономическая (физическая, социальная и т.п.) сущность исследуемого явления или процесса, иначе в большинстве случаев будет получаться, что наилучшим приближающим уравнением является полином 6-й степени.

Для рассматриваемой задачи уравнения, которые отвечают ее экономической сущности, имеют линейный или степенной вид. Для того чтобы получить аналитическое выражение выбранного уравнения, необходимо на вкладке ***Параметры*** (*см.* рис. 16.2) акти-

23-2817

визировать флажок *Показывать уравнение на диаграмме*. Если активизировать флажок *Поместить на диаграмму величину достоверности аппроксимации R^2*, то в области построения будет выведено значение показателя R^2, по которому можно судить, насколько хорошо выбранное уравнение аппроксимирует эмпирические данные. Чем ближе R^2 к единице, тем уравнение является более адекватным исследуемому явлению или процессу.

На рис. П.2 изображен график линейной зависимости спроса от цены с выводом в области построения аналитического выражения уравнения и значения показателя R^2.

Итак, мы получили, что *Спрос* = –456,65 × Цена + 15610. Подставляя данное уравнение в уравнение прибыли, получим ее теоретические (расчетные) значения (табл. П.12, диапазон C10:H10)

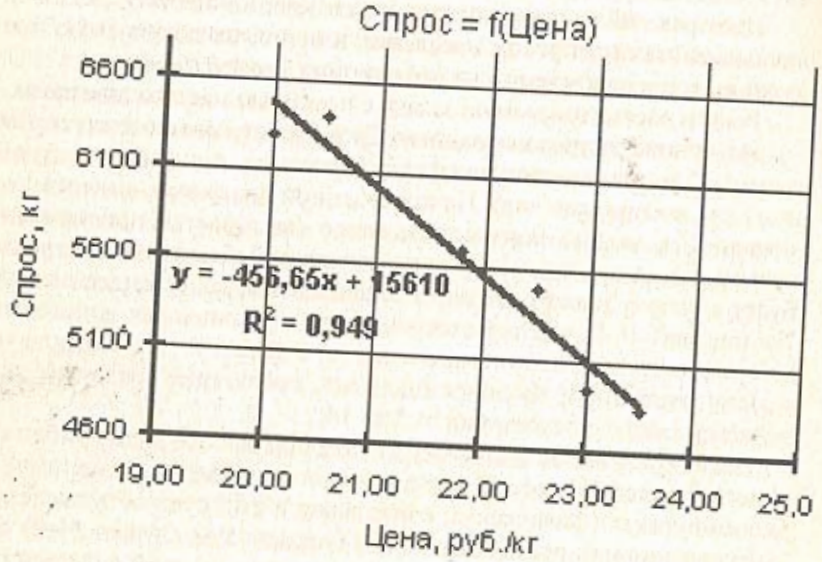

Рис. П.2

Ну а как же найти максимальную прибыль, и при какой цене и объемах производства она достигается? И здесь нам опять может прийти на помощь мастер диаграмм.

Таблица П.12

	B	C	D	E	F	G	H
4		январь	февраль	март	апрель	май	июнь
5	Цена, руб./кг	20,00	20,50	23,50	23,00	22,50	21,80
6	Продано, кг	6300	6400	4850	4950,00	5500,00	5700,00
7	Общая выручка, руб.	126000,00	131200,00	113975,00	113850,00	123750,00	124260,00
8	Общие издержки, руб.	95600,00	96800,00	78200,00	79400,00	86000,00	88400,00
9	Прибыль, руб.	30400,00	34400,00	35775,00	34450,00	37750,00	35860,00
10	Прибыль, руб. (теоретическая)	31816,00	33113,74	36105,34	36177,55	36021,44	35419,29

Легко заметить, что после подстановки уравнения спроса в уравнение прибыли последнее приобретает квадратичный вид, т.е. является параболой, или, иными словами, полиномом 2-й степени. Принимая это во внимание, на основе диапазона данных C5:H5 и C10:H10 строим точечную диаграмму, значения которой аппроксимируем полиномом 2-й степени по рассмотренной выше технологии. Результат аппроксимации изображен на рис. П.3.

Как видно из рис. П.3, максимальная прибыль $Pr^* \approx 36200$ руб. достигает при цене $P^* \approx 23$ руб./кг. Подставляя значение 23 руб./кг в уравнение спроса, получаем оптимальный объем производства $Q^* \approx 5107$ кг.

Итак, решение найдено. Главным его преимуществом является простота и наглядность, главным недостатком – некоторая неточность, присущая, впрочем, практически каждому графическому решению. Уменьшить эту неточность можно путем изменения максимальных и минимальных значений шкал диаграммы (рис. П. 4).

Если вы все же не являетесь сторонником графических решений и для нахождения экстремумов предпочитаете вычислять производные, то для автоматизации расчетов целесообразно разработать соответствующие пользовательские функции.

После анализа расчетов, проведенных в рассмотренной задаче, были разработаны две пользовательские функции – для случаев линейной и степенной зависимостей между спросом и ценой.

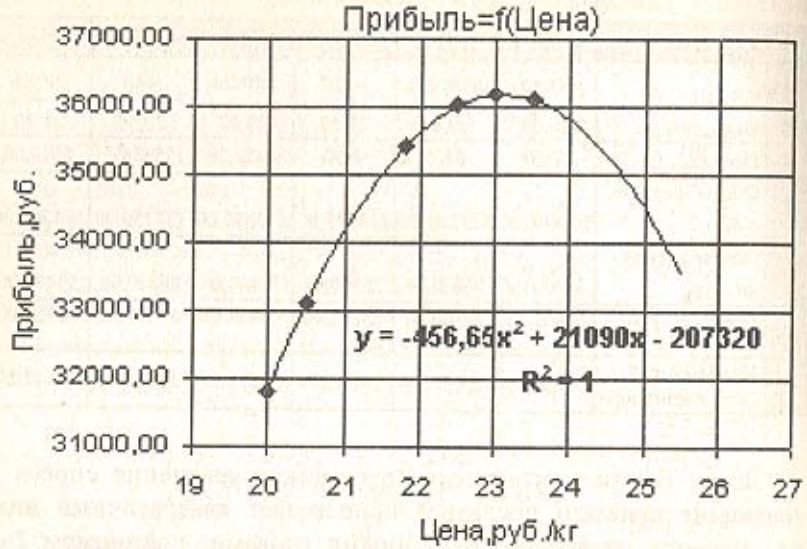

Рис. П.3

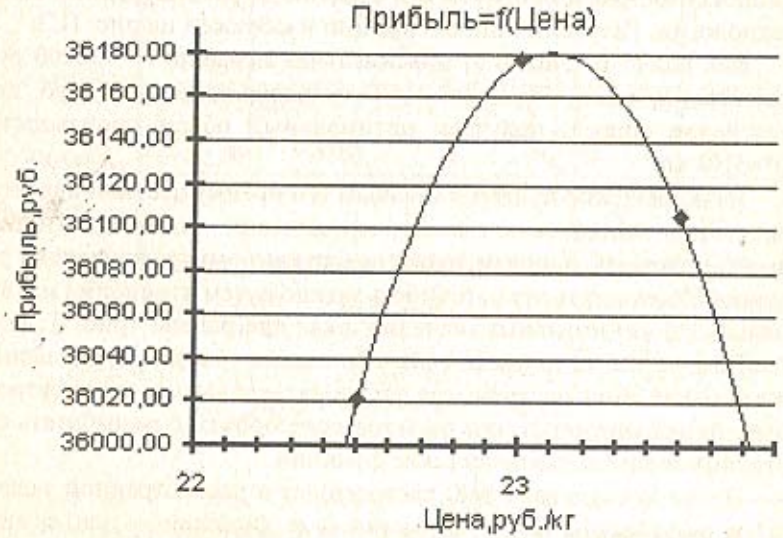

Рис. П.4

Текст программы для линейной зависимости

```
Function Линейная(Коэффициент As Single, Свободный_член As
Single, _
        AVC As Single, Optional FC As Single)
    Dim Цена As Single, Объем As Single, Прибыль As Variant
    Application.Volatile True
    Цена = (AVC * Коэффициент - Свободный_член) / (2 * Коэф-
фициент)
    Объем = Коэффициент * Цена + Свободный_член
    If FC = 0 Then
      Прибыль = «введите FC»
    Else
      Прибыль = Цена * Объем - FC - AVC * Объем
    End If
    Линейная = Array(Array(«Цена», «Объем», «Прибыль»), _
               Array(Цена, Объем, Прибыль))
End Function
```

Текст программы для степенной зависимости

```
Function Гипербола(Коэффициент As Single, Показатель_степе-
ни As Single, _
        AVC As Single, Optional FC As Single)
    Dim Цена As Single, Объем As Single, Прибыль As Variant
    Application.Volatile True
    Цена = (AVC * Показатель_степени) / (Показатель_степени + 1)
    Объем = Коэффициент * Цена ^ Показатель_степени
    If FC = 0 Then
      Прибыль = «введите FC»
    Else
      Прибыль = Цена * Объем - FC - AVC * Объем
    End If
    Гипербола = Array(Array(«Цена», «Объем», «Прибыль»), _
                Array(Цена, Объем, Прибыль))
End Function
```

В качестве аргументов функций используются параметры приближающих уравнений, а также значения постоянных (*FC*) и сред-

них переменных издержек (*AVC*). Постоянные издержки (*FC*) являются необязательным аргументом, поэтому они могут быть опущены при задании функций, но в этом случае не будет рассчитываться значение прибыли.

Примечание. Для расчета параметров уравнения линейной регрессии можно использовать функцию ЛИНЕЙН, для расчета параметров уравнения экспоненциальной (показательной) регрессии – функцию ЛГРФПРИБЛ.

Так как разработанные функции вычисляют несколько величин (цену, объем производства и прибыль), то их следует вводить как формулы массива, т. е. выделить диапазон ячеек, куда будут помещены расчетные значения, ввести функцию и нажать комбинацию клавиш **Ctrl+Shift+Enter**. Введенная функция будет автоматически заключена в фигурные скобки {}. В разработанных функциях значения рассчитываются в массиве размерностью 2×3, причем первая строка является «шапкой» таблицы. Используя в качестве аргументов функции ЛИНЕЙНАЯ параметры приближающего уравнения и значения *AVC* и *FC* рассмотренной задачи, получим результаты, представленные в табл. П.13.

Таблица П.13

	C	D	E
33	Цена	Объем	Прибыль
34	23,09	5065,10	36181,40

Если вы предполагаете использовать разработанные вами функции и в других рабочих книгах (или на других компьютерах), то рабочую книгу, которой они принадлежат, целесообразно сохранить с расширением *xla*. В этом случае рабочая книга будет сохранена как надстройка Microsoft Excel, которая может быть впоследствии перенесена на любой компьютер и подключена к Excel через окно **Надстройки** (*см.* рис. 1.2).

УКАЗАТЕЛЬ СТАТИСТИЧЕСКИХ ФУНКЦИЙ

ЛИТЕРАТУРА

1. *Вентцель Е. С.* Теория вероятностей. – М.: Наука, 1964. – 577 с.

2. *Доклад* Ярославского областного комитета государственной статистики за 1998 год. – Ярославль.: ЯОКГС, 1999. – 199 с.

3. *Замков О. О., Толстопятенко А. В., Черемных Ю. Н.* Математические методы в экономике. – М.: МГУ, Изд-во ДИС, 1997. – 367 с.

4. *Ильина О. П., Макарова Н. В.* Статистический анализ и прогнозирование экономической информации в электронной таблице Excel 5.0 Microsoft: Учеб. пособие. – СПб.: Изд-во СПбУЭФ, 1996. – 140 с.

5. *Калинина В. Н., Панкин В. Ф.* Математическая статистика. – М.: Высш. шк., 1998. – 336 с.

6. *Кимбл Г.* Как правильно пользоваться статистикой. – М.: Финансы и статистика, 1969. – 294 с.

7. *Математический* энциклопедический словарь/Под ред. Ю. В. Прохорова. – М.: Сов. энциклопедия, 1988. – 847 с.

8. *Общая* теория статистики: Статистическая методология в изучении коммерческой деятельности: Учебник/Под ред. О. Э. Башиной, А. А. Спирина. – М.: Финансы и статистика, 2000. – 439 с.

9. *Общая* теория статистики: Учебник/Г. С. Кильдишев, В. Е. Овсиенко и др. – М.: Статистика, 1980. – 423 с.

10. *Справочник* по вероятностным расчетам/Г. Г. Абезгауз, А. П. Тронь и др. – М.: Военное изд-во, 1970. – 537 с.

11. *Справочник* по прикладной статистике. В 2-х т.: Пер. с англ./Под ред. Э. Ллойда, У. Ледермана, Ю. Н. Тюрина. – М.: Финансы и статистика, 1989. – 510 с.

12. *Теория* статистики: Учебник/Под ред. Р. А. Шмойловой. – М.: Финансы и статистика, 2001. – 557 с.

13. *Тюрин Ю. Н., Макаров А. А.* Статистический анализ данных на компьютере/Под ред. В. Э. Фигурнова. – М.: Инфра-М, 1998. – 528 с.

14. *Хан Г., Шапиро С.* Статистические модели в инженерных задачах. – М.: Мир, 1969. – 395 с.

ОГЛАВЛЕНИЕ

Раздел II. МЕТОДЫ ПРОВЕРКИ СТАТИСТИЧЕСКИХ ГИПОТЕЗ

Раздел III. ДИСПЕРСИОННЫЙ АНАЛИЗ

Раздел IV. СТАТИСТИЧЕСКИЕ МЕТОДЫ ИЗУЧЕНИЯ ВЗАИМОСВЯЗЕЙ ЯВЛЕНИЙ И ПРОЦЕССОВ

Раздел V. СТАТИСТИЧЕСКИЕ МЕТОДЫ ИЗУЧЕНИЯ ДИНАМИКИ ПРОЦЕССОВ

Учебное издание

Макарова Наталья Владимировна
Трофимец Валерий Ярославович

СТАТИСТИКА В EXCEL

Заведующая редакцией *Л. А. Табакова*
Редактор *А. М. Маторина*
Младший редактор *Н. А. Федорова*
Художественный редактор *Ю. И. Артюхов*
Технический редактор *Т. С. Маринина*
Корректоры *Н. Б. Вторушина, Г. В. Хлопцева*
Компьютерная верстка *Е. А. Жигунова, Е. А. Бычинская*
Оформление художника *Н. М. Биксентеева*

ИБ № 4179

Подписáно в печать 06.08.2003. Формат 60×88 $^{1}/_{16}$. Печать офсетная
Гарнитура «Таймс». Усл. п. л. 22,54. Уч.-изд. л. 19,9
Тираж 4000 экз. Заказ 2817. «С» 200

Издательство «Финансы и статистика»
101000, Москва, ул. Покровка, 7
Телефон (095) 925-35-02, факс (095) 925-09-57
E-mail: mail@finstat.ru http://www.finstat.ru

ГУП «Великолукская городская типография»
Комитета по средствам массовой информации Псковской области,
182100, Великие Луки, ул. Полиграфистов, 78/12
Тел./факс: (811-53) 3-62-95
E-mail: VTL@MART.RU