



## Глава 7

# Таблицы

В этой главе...



|  |     |
|--|-----|
| ▶ Ñaī aī Ūá òàáèèöŪ  | 262 |
| ▶ ÒàáèèöŪ īī āāōī ēàòāāī ðèàèüī Ūī ī āđāī āī í Ūī  | 271 |
| ▶ ĀŪ÷èñēāī ēā ī æèāāāī ī āī ēī èè÷āñòāā ī àáèpāāī èé                                       | 274 |
| ▶ Ñòàòèñòèēā òè-éāāāđàò ī èđñī ī a   | 276 |
| ▶ Āđóāēā ñòàòèñòè÷āñēēā ī āđāī áòđŪ òàáèè÷ī Ūō āāī í Ūō                                    | 280 |
| ▶ Ī āī ñī ī āāī ī ī ñòū ī ðèī āī āī èy òāñòà òè-éāāāđàò ī èđñī ī a<br>āèy ī àèŪō āŪāī đī ē | 282 |
| ▶ ÒàáèèöŪ ñ ī ī đyāēī āŪī è ī āđāī āī í Ūī è   | 285 |
| ▶ Óī ðāæī āī èy  | 290 |



**В** этой главе рассматриваются способы обработки категориальных данных в виде таблиц и порядковых переменных. Здесь описывается мастер программы Excel для создания сводных таблиц, а также инструменты модуля StatPlus для анализа данных в таблицах.

## Сводные таблицы

В предыдущей главе описано использование t-тестов и непараметрических тестов для анализа непрерывных переменных. Для анализа категориальных и порядковых данных также можно применять описанные ранее способы проверки гипотез. Данные такого типа часто встречаются в опросных листах, которые разбиты на категории. Допустим, у нас есть таблица со списком преподавателей, разбитым на категории по должности (доцент, адъюнкт-профессор или профессор) и полу (мужской или женский). Можно ли с достаточной статистической значимостью утверждать, что профессоров-мужчин больше, чем профессоров-женщин? Сколько профессоров-женщин могло бы быть при отсутствии дискриминации по половому признаку? Анализ категориальных данных позволяет дать ответ на эти и многие другие вопросы.

Для иллюстрации способов работы с категориальными данными рассмотрим данные, полученные после опроса преподавателей статистики и собранные в рабочей книге `Опрос.xls` (`Survey.xls`). В этом файле находятся данные из 392 опросных листов с вопросами о предварительном знании основ исчисления, необходимости программного обеспечения, типе используемого компьютера и т.д. В табл. 7.1 приводится краткое описание используемых переменных.

**Таблица 7.1. Структура результатов опроса преподавателей статистики**

| Имя диапазона                      | Диапазон | Описание   |
|------------------------------------|----------|--|
| Компьютер (Computer)               | A2:A393  | Тип используемого компьютера   |
| Кафедра (Dept)                     | B2:B393  | Кафедра  |
| Доступность (Available)            | C2:C393  | Тип компьютерной системы, доступной студентам  |
| Интерес (Interest)                 | D2:D393  | Степень интереса к дополнительным статистическим материалам  |
| Исчисление (Calculus)              | E2:E393  | Необходимость знания основ исчисления  |
| Исп_прогр_обесп<br>(Uses_Software) | F2:F393  | Необходимость использования программного обеспечения   |
| Зачислено_A (Enroll_A)             | G2:G393  | Категориальная переменная для обозначения семестрового количества студентов. Например, 001-050 означает, что в течение семестра обучаются от 1 до 50 студентов |
| Зачислено_B (Enroll_B)             | H2:H393  | Категориальная переменная для обозначения годового количества студентов  |
| Макс_Стоимость (Max_Cost)          | I2:I393  | Максимальная стоимость вспомогательных компьютерных материалов   |

×õíáÙ îðèðÙòù ðááí÷òð éí èáó `Опрос.xls` (`Survey.xls`), áÙíîéí èðá í áðá÷èñéáí í Ùá í èæá ááéñòáèý.

1. Запустите программу Excel и разверните ее окно на весь экран.
2. Откройте рабочую книгу `Опрос.xls` (`Survey.xls`), которая находится в каталоге `Примеры` (`Student`).
3. Сохраните рабочую книгу в файле `Опрос2.xls` (`Survey2.xls`) (рис. 7.1).

|    | A         | B         | C           | D          | E            | F                       | G            | H            |
|----|-----------|-----------|-------------|------------|--------------|-------------------------|--------------|--------------|
| 1  | Компьютер | Кафедра   | Доступность | Интерес    | Исчисление   | Использование программы | Зачисление_A | Зачисление_B |
| 2  | ПК        | Матем     | Лаборатория | высокий    | Не требуется | да                      | 001-050      | 051          |
| 3  | ПК        | Матем     | Мэйнфрейм   | высокий    | Требуется    | да                      |              |              |
| 4  | ПК        | Матем     | Мэйнфрейм   | наибольший | Не требуется | да                      | 101-150      | 101          |
| 5  |           |           | Другое      | наибольший | Не требуется | да                      | 001-050      | 001          |
| 6  |           |           | Лаборатория | векторный  | Требуется    | да                      |              |              |
| 7  | Мэйнфрейм |           | Другое      | высокий    | да           |                         |              |              |
| 8  |           |           |             | векторный  | Не требуется | может быть              |              |              |
| 9  | ПК        |           | Лаборатория | наименьший | Не требуется | да                      |              |              |
| 10 |           |           | Другое      | наибольший | да           |                         | 201-300      | 301          |
| 11 | ПК        |           | Другое      | векторный  | Не требуется | да                      |              |              |
| 12 | ПК        |           | Лаборатория | наибольший | Не требуется | да                      |              |              |
| 13 | Мэйнфрейм |           | Мэйнфрейм   | наименьший | Не требуется | да                      |              |              |
| 14 | ПК        | Бис_Экон  | Мэйнфрейм   | высокий    | Требуется    | да                      | 051-100      | 101          |
| 15 |           |           | Другое      | наибольший | Не требуется | да                      |              |              |
| 16 | ПК        |           | Лаборатория | векторный  | Не требуется | да                      |              |              |
| 17 | ПК        |           | Другое      | векторный  | Не требуется | да                      |              |              |
| 18 |           | Матем     |             |            | Требуется    | нет                     | 001-050      | 001          |
| 19 | Мэйнфрейм | Матем     | Мэйнфрейм   | наименьший | Не требуется | да                      | 001-050      | 001          |
| 20 | Мэйнфрейм |           | Мэйнфрейм   | малый      | Не требуется | да                      | 051-100      |              |
| 21 | ПК        | Мед_Здрав | Лаборатория | векторный  | Не требуется | да                      | 101-150      | 151          |
| 22 |           | Социал    |             | высокий    | Не требуется | может быть              | 001-050      | 001          |
| 23 | ПК        | Бис_Экон  | Другое      | высокий    | Не требуется | да                      | 501-         | 501          |
| 24 |           | Бис_Экон  |             | векторный  | Не требуется | может быть              | 751-100      | 471          |

Рис. 7.1. Таблица исходных данных

Допустим, что требуется определить, какие компьютеры наиболее часто упоминаются в данных опроса в столбце Компьютер (Computer). Возможно четыре варианта ответа: ПК (PC), Mac (Mac), Мини-компьютер (Minicomputer) и Мэйнфрейм (Main).

## Программа-мастер для создания сводных таблиц

В Excel для подсчета количества ответов по каждому варианту предусмотрена *pivot table* (результатов опроса в этом примере). Сводные таблицы Excel являются интерактивными, т.е. могут автоматически обновляться при изменении исходных данных. Кроме того, для просмотра данных разными способами можно поворачивать таблицу для отображения данных по строкам или по столбцам. (Именно поэтому в английское название сводной таблицы входит слово *pivot*, т.е. поворачивать.)

Попробуйте создать сводную таблицу, в которой подытоживаются данные опроса преподавателей об используемом компьютере. Для этого следует применить программу — мастер сводных таблиц Excel аналогично мастеру диаграмм Excel, что существенно упрощает процесс создания сводных таблиц.

Шаги создания сводной таблицы:

1. Выберите команду меню Данные⇒Сводная таблица (Data⇒PivotTable and PivotChart Report), чтобы открыть первое диалоговое окно мастера сводных таблиц и диаграмм.
2. Установите переключатель в списке или базе данных Microsoft Excel (Microsoft Excel List or Database) в верхней группе переключателей и переключатель сводная таблица (PivotTable) в нижней группе переключателей.

3. Щелкните на кнопке **Далее**.

На экране появится второе диалоговое окно мастера сводных таблиц и диаграмм, в котором необходимо указать диапазон исследуемых данных в поле **Диапазон (Range)**.

4. Укажите значение **\$A\$1:\$I\$393** в поле **Диапазон (Range)**, а затем щелкните на кнопке **Далее** для перехода к третьему диалоговому окну мастера сводных таблиц.

Теперь можно приступить к оформлению внешнего вида сводной таблицы. Сводная таблица имеет четыре области. В области **Строка (Row)** располагаются категории, которые отображаются в строках таблицы, а в области **Столбец (Column)** — категории, которые отображаются в столбцах таблицы. В области **Данные (Data)** располагаются значения, которые отображаются на пересечениях строк и столбцов. Наконец, область **Страница (Page)** используется для создания разных таблиц для разнообразных категорий категориальной переменной.

Àëü ñî çääíëü äí äøí äâí äëää ñâí äíí é èääëëüü äüííëí èðâ í äðâ-ëñëâí í üâ í èæâ ääëñðäëü.

1. В третьем диалоговом окне мастера сводных таблиц щелкните на кнопке **Макет (Layout)**.

На экране появится схематическое представление макета сводной таблицы. В данном окне можно по своему усмотрению располагать разные переменные таблицы, перетаскивая расположенные справа кнопки полей.

2. Перетащите кнопку **Компьютер (Computer)** в область **Строка (Row)**.

Это значит, что Excel должна использовать в строках варианты компьютеров ПК (PC), Mac (Mac), Мини-компьютер (Minicomputer) и Мэйнфрейм (Main).

3. Перетащите кнопку **Компьютер (Computer)** в область **Данные (Data)**.

После этого в области **Данные (Data)** появится подпись **Количество** по полю **Компьютер (Count of Computer)**, что означает отображение в сводной таблице количества ответов с указанием каждого варианта компьютера. После выполнения этих действий диалоговое окно макета сводной таблицы будет выглядеть так, как на рис. 7.2.

4. Щелкните на кнопке **OK**.

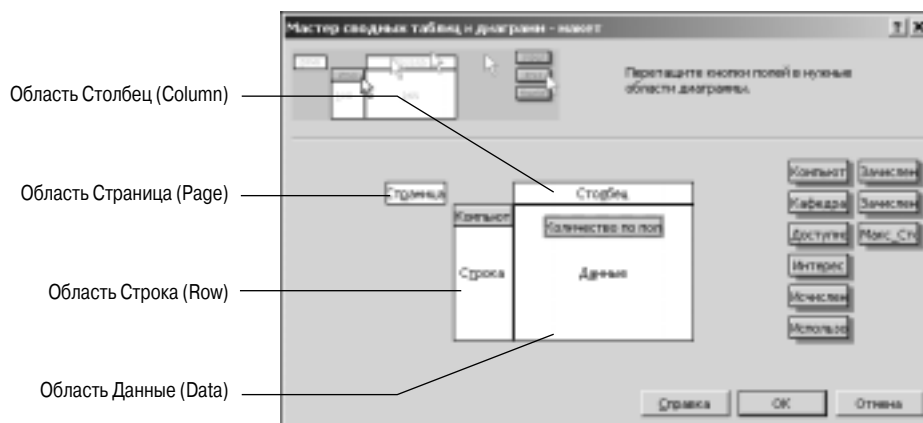


Рис. 7.2. 1 äëâð ñâí äíí é èääëëüü

Теперь можно приступить к редактированию внешнего вида сводной таблицы.

Àëý òí ðààèáí èý ïàðàì àòðàì è áí àðíááí àèàà ñáíáííé ðààèèòù áùííéí èðà ïáðà-èñèáí-í ùá í èæá àáèñàèý.

1. В третьем диалоговом окне мастера сводных таблиц щелкните на кнопке Параметры (Options).
2. В появившемся диалоговом окне Параметры сводной таблицы (PivotTable Options), которое показано на рис. 7.3, снимите флажок Общая сумма по строкам (Grand totals for rows).

Итоговые значения для строк позволяют вычислять сумму количеств наблюдений для разных уровней категориальной переменной. В данном примере в области Столбец (Column) нет категориальной переменной, поэтому не нужно вычислять итоговые значения для строк.

3. Введите значение 0 в текстовом поле для пустых ячеек отображать (For empty cells, show).
4. Это значит, что при отсутствии наблюдений в отдельных ячейках таблицы программа Excel автоматически отобразит в них значение 0 (см. рис. 7.3).

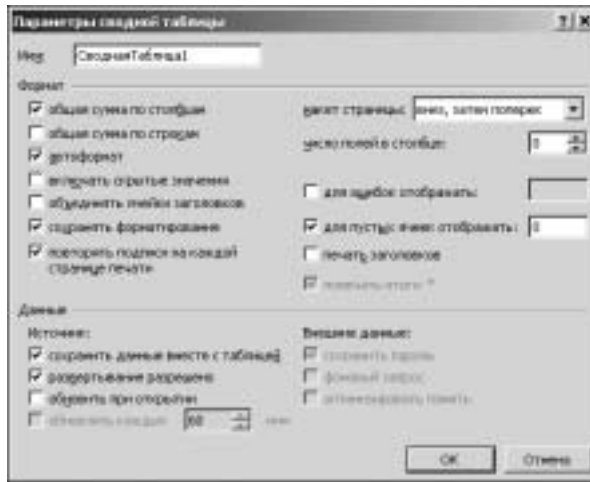


Рис. 7.3. Параметры сводной таблицы

5. Щелкните на кнопке ОК.
6. В третьем диалоговом окне мастера сводных таблиц выберите переключатель новый лист (New worksheet), а затем щелкните на кнопке Готово.

На рис. 7.4 показан окончательный вид созданной сводной таблицы.

В первом столбце сводной таблицы перечислены четыре категории переменной Компьютер (Computer) — ПК (PC), Mac (Mac), Мини-компьютер (Minicomputer), Мэйнфрейм (Main) — и пустая ячейка (A3) для отсутствующих значений. Во втором столбце приведено количество ответов преподавателей с указанием каждого типа компьютера, который используется в процессе обучения. (Программа Excel может также отобразить панель инструментов для редактирования сводной таблицы. В данной главе эта панель инструментов не используется, однако более подробные сведения о ней можно получить в контекстной оперативной справке Excel.) Как видите, 130 из 392 преподавателей используют ПК, 57 — мэйнфреймы, 24 — компьютеры Macintosh, 4 —

мини-компьютеры, а 177 не дали ответа или их ответ не зафиксирован. Количество отсутствующих данных может быть очень большим из-за того, что в некоторых учебных курсах используются разные виды компьютеров. Обычно в таких курсах чаще других видов компьютеров используются ПК.

| Количество по полю Компьютер | Итого |
|------------------------------|-------|
| Компьютер                    | 177   |
| Mac                          | 24    |
| Мини-компьютер               | 4     |
| Мейнфрейм                    | 57    |
| ПК                           | 130   |
| Общий итог                   | 352   |

Рис. 7.4. Παίξιμο ελεγχών πηγαίων δεδομένων και εμφάνιση των αποτελεσμάτων σε πίνακα

## Удаление категорий из сводной таблицы

В данном примере нас не интересует количество отсутствующих данных, а потому нужно удалить эту категорию из сводной таблицы. Сводная таблица содержит разворачивающийся список флажков, которые используются для указания отображаемых в таблице категорий.

Αυτή η διαδικασία είναι απαραίτητη εάν θέλετε να αφαιρέσετε κάποιες κατηγορίες.

1. Щелкните на стрелке возле поля Компьютер (Computer) в сводной таблице.
2. В развернувшемся списке снимите флажок параметра без подписи, как показано на рис. 7.5.



Рис. 7.5. Αφαίρεση της κατηγορίας από τον κατάλογο

3. Щелкните на кнопке ОК.

После этого сводная таблица изменится и не будет содержать категорию для отсутствующих данных (рис. 7.6).

Теперь сводная таблица содержит ответы 215 преподавателей, которые указали тип используемого компьютера; 130 из 215 ответов относятся к персональному компьютеру. Для продолжения анализа можно удалить другие категории, например для сравнения количества персональных компьютеров и компьютеров типа Macintosh.

|   | А                             | В     |
|---|-------------------------------|-------|
| 1 | Генерировать сводную таблицу  |       |
| 2 |                               |       |
| 3 | Количество по полю Контакттер |       |
| 4 | Контакттер                    | Итого |
| 5 | Мас                           | 24    |
| 6 | Мини-компьютер                | 4     |
| 7 | Мэйнфрейм                     | 57    |
| 8 | ПК                            | 130   |
| 9 | Общий итог                    | 215   |

Рис. 7.6. Настройка сводной таблицы

## Изменение значений в сводной таблице

По умолчанию сводная таблица отображает количество наблюдений каждой категории в результатах опроса. Однако в ней можно отображать многие другие типы значений: суммы, максимумы, минимумы, средние и процентные доли. Например, при отображении процентных долей в таблице указываются данные для всей таблицы, строки или столбца. Попробуйте изменить вид сводной таблицы для отображения процентной доли наблюдений в каждой категории.

Для изменения значений в сводной таблице выполните следующие действия:

- Щелкните правой кнопкой мыши на одной из ячеек в столбце В с количеством наблюдений, а затем выберите команду Параметры поля (Field Settings) в контекстном меню.
- В диалоговом окне Вычисление поля сводной таблицы щелкните на кнопке Дополнительно>> (Options>>) для разворачивания нижней части окна.
- В списке Дополнительные вычисления (Show Data As) выберите элемент Доля от суммы по столбцу (% of column) и щелкните на кнопке ОК.

После выполнения этих действий сводная таблица изменится и в столбце В будут отображены процентные доли каждой категории от суммы по столбцу (рис. 7.7).

|   | А                             | В       |
|---|-------------------------------|---------|
| 1 |                               |         |
| 2 |                               |         |
| 3 | Количество по полю Контакттер |         |
| 4 | Контакттер                    | Итого   |
| 5 | Мас                           | 11,16%  |
| 6 | Мини-компьютер                | 1,86%   |
| 7 | Мэйнфрейм                     | 26,51%  |
| 8 | ПК                            | 60,47%  |
| 9 | Общий итог                    | 100,00% |

Рис. 7.7. Настройка сводной таблицы

Как видите, 60,47% давших ответ преподавателей используют в своих учебных курсах персональные компьютеры, 26,51% — мэйнфреймы, 11,16% — компьютеры Macintosh, а 2% — мини-компьютеры.

Для изменения значений в сводной таблице выполните следующие действия:

- Щелкните правой кнопкой мыши на одной из ячеек в столбце В с процентной долей, а затем выберите команду Параметры поля в контекстном меню.

2. В списке Дополнительные вычисления выберите элемент Нет (Normal) и щелкните на кнопке ОК.

## Отображение наблюдений в столбиковой диаграмме

Данные из сводной таблицы можно отображать в сводной диаграмме. По умолчанию сводная диаграмма имеет вид столбиковой диаграммы, в которой высота столбика пропорциональна количеству наблюдений в каждой ячейке.

×ōīáŭ ñīçááōŭ ñōīéáéēīáōŕ äēááðáī ì ó äëŷ îōīáðáæáíëŷ äáííŭō íá ëñííëŷçóáí ŭō ëíí ïŭŕóðáðō èç ñáíáííé ðááéëōŭ, áŭííëíëðá ðéáçáííá íëæá äáéñóáëá.

1. Щелкните правой кнопкой мыши внутри сводной таблицы и выберите команду Сводная диаграмма (PivotChart) в контекстном меню.



Àëÿ àèèþ=àí èÿ òòàíèíàé ñ òèàçàí èàí èí èè=àíðàà í ààèþààí èè àùííèí èðà ò àðà=èíèàí í ùà í èàà ààéíðàèÿ.

1. Выберите команду меню Диаграмма⇒Параметры диаграммы (Chart⇒Chart Options).
2. Выберите вкладку Подписи данных (Data Labels).
3. Установите флажок значения (Value) и щелкните на кнопке ОК.

После этого Excel обновит диаграмму, в которой теперь будут отображены надписи с указанием количества ответов для каждой категории (рис. 7.9).

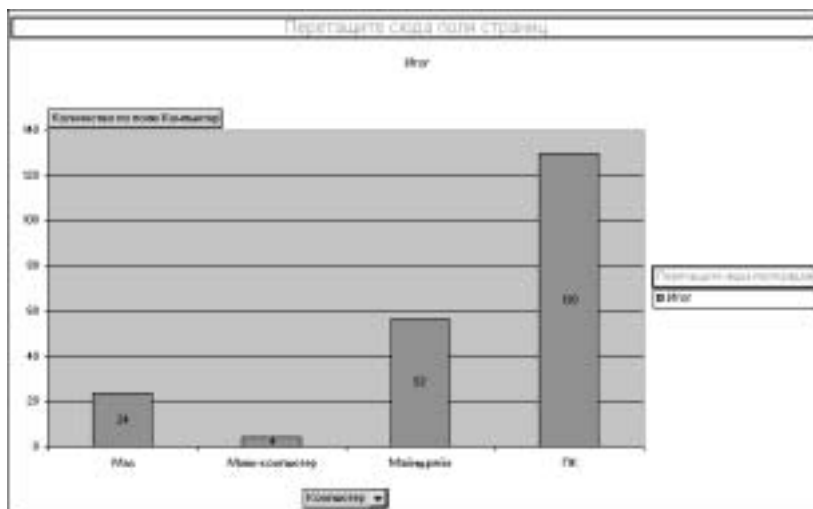


Рис. 7.9. Количество ответов на вопросы по использованию ПК и мобильных устройств

На рис. 7.9 видно, что персональные компьютеры используются в 130 учебных курсах, и с помощью несложных арифметических действий можно установить, что это составляет больше половины всех учебных курсов.

## Отображение категориальных данных в круговой диаграмме

Еще один способ сравнения размеров отдельных групп с их суммой основан на использовании круговой диаграммы. Круговая диаграмма (или пирог — pie), в котором каждый сектор (или кусочек пирога) представляет свою категорию. Попробуем применить круговую диаграмму для представления данных о компьютерах. Вместо создания круговой диаграммы с самого начала попробуем преобразовать имеющуюся столбиковую диаграмму в круговую.

Àëÿ òòàíèíàçíàí èÿ òòàíèíàé ñ àààðàí ò à èðàíàòþ àùííèí èðà ò àðà=èíèàí í ùà í èàà ààéíðàèÿ.

1. Выберите команду меню Диаграмма⇒Тип диаграммы (Chart⇒Chart Type).
2. В списке Тип (Chart types) выберите элемент Круговая (Pie) и щелкните на кнопке ОК.

После выполнения этих действий столбиковая диаграмма превратится в круговую (рис. 7.10).

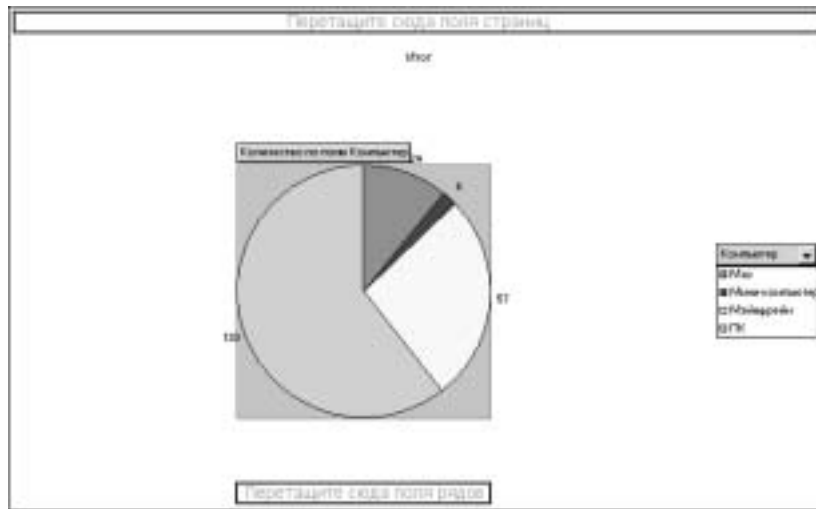


Рис. 7.10. Использование средств Power Point

С помощью этой круговой диаграммы можно легко определить, что персональные компьютеры составляют больше половины используемых компьютеров. С помощью столбиковой диаграммы эту информацию можно было получить только после выполнения арифметических действий с количеством наблюдений в каждой категории, которые указаны в подписях. Круговая диаграмма по-прежнему содержит эти подписи, но их можно удалить или видоизменить. Попробуем отобразить в подписях процентные доли каждой категории и имена категорий.

Для этого выполните следующие действия:

1. Щелкните правой кнопкой мыши на круговой диаграмме и выберите команду Формат рядов данных (Format Data Series).
2. Выберите вкладку Подписи данных.
3. Во вкладке Подписи данных установите флажки имена категорий (Category Name) и доли (Percentage).
4. Щелкните на кнопке ОК.

После выполнения этих действий круговая диаграмма будет выглядеть так, как на рис. 7.11.

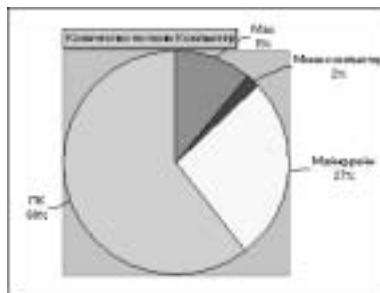


Рис. 7.11. Использование средств Power Point

Круговая диаграмма очень эффективна для отображения категориальных данных, поэтому ее чаще используют в бизнес-отчетах, чем в статистических отчетах.



#### Советы

- Для просмотра исходных данных в ячейке с количеством наблюдений любой категории сводной таблицы дважды щелкните на ней. После этого Excel автоматически откроет новый лист с отображением всех наблюдений данной категории.
- В качестве источника данных сводной таблицы, помимо текущей рабочей книги, можно использовать базы данных и внешние файлы.
- При изменении значений данных в источнике данных сводная таблица автоматически обновляется для отражения внесенных изменений.

---

## Таблицы по двум категориальным переменным

Допустим, требуется исследовать взаимосвязь между двумя категориальными переменными, например использование компьютеров на разных кафедрах. Верно ли, что на разных кафедрах разные компьютеры используются в разных пропорциях? Верно ли, что на одной кафедре предпочитают использовать компьютеры Macintosh, а на другой — персональные компьютеры? В Excel предусмотрена возможность создания сводной таблицы для количества наблюдений разных компьютеров по кафедрам. В данном примере упоминаются следующие кафедры:

- Биз\_Экон (Bus ,Econ) — кафедра бизнеса и экономики;
- Мед\_Здрав (HealthSci) — кафедра медицины и здравоохранения;
- Матем (MathSci) — кафедра математики, статистики, физики и техники;
- Социол (SocSci) — кафедра социологии.

Аёу пī с̄ааіёу п̄аі аі іё оааёёоᵀ п̄ оёасаіёаі ёіёё-аᵀоаа ёіі іᵀроаᵀі а̄ дасі оо аеаіа іі ёаᵀааᵀаі аᵀіі іёі ёᵀа і̄аᵀа-ёᵀёаі іᵀа іёаа̄ ааёᵀᵀаёу.

1. Вернитесь к рабочему листу Результаты опроса (Survey).
2. Выберите команду меню Данные⇒Сводная таблица (Data⇒PivotTable and PivotChart Report).  
При использовании того же источника данных, что и для предыдущей сводной таблицы, ее не нужно повторно выделять.
3. Выберите переключатель в другой сводной таблице или диаграмме (Another PivotTable or PivotChart Report), а затем щелкните на кнопке Далее.  
В Excel отображается список сводных таблиц из данной книги, и, поскольку в этом примере есть только одна сводная таблица, именно она и будет выбрана автоматически.
4. Щелкните на кнопке Далее.
5. В третьем диалоговом окне мастера сводных таблиц и диаграмм щелкните на кнопке Макет.
6. Перетащите кнопку Компьютер (Computer) в область Строка (Row), а затем в область Данные (Data).
7. Перетащите кнопку Кафедра (Dept) в область Столбец (Column), как показано на рис. 7.12.
8. Щелкните на кнопке ОК.
9. В третьем диалоговом окне мастера сводных таблиц и диаграмм щелкните на кнопке Параметры.

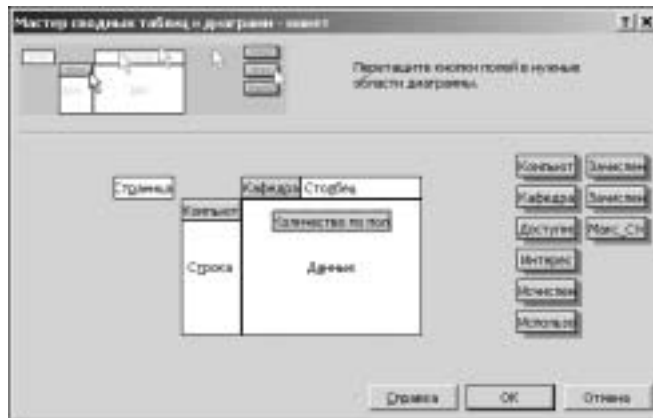


Рис. 7.12. Настраиваем параметры сводной таблицы

10. Введите значение 0 в текстовом поле для пустых ячеек отображать.
11. Щелкните на кнопке Готово.

Как и при создании предыдущей сводной таблицы попробуем скрыть категорию для отсутствующих данных.

Для этого выполните следующие действия:

1. Щелкните на стрелке справа возле поля Компьютер (Computer) в сводной таблице и в развернувшемся списке снимите флажок параметра без подписи. Щелкните на кнопке ОК.
2. Щелкните на стрелке справа возле поля Кафедра (Dept) в сводной таблице и в развернувшемся списке снимите флажок параметра без подписи. Щелкните на кнопке ОК.

На рис. 7.13 показан окончательный вид сводной таблицы.

|    | A                            | B                                  | C     | D           | E      | F          | G |
|----|------------------------------|------------------------------------|-------|-------------|--------|------------|---|
| 1  |                              | (протяните сюда, если есть строки) |       |             |        |            |   |
| 2  |                              |                                    |       |             |        |            |   |
| 3  | Количество по полю Компьютер | Кафедра                            |       |             |        |            |   |
| 4  | Компьютер                    | Созд. Экон                         | Матем | Мед. Социол | Социол | Общий итог |   |
| 5  | Mac                          | 4                                  | 12    | 0           | 6      | 22         |   |
| 6  | Мини-компьютер               | 1                                  | 2     | 0           | 1      | 4          |   |
| 7  | Майнфрейм                    | 12                                 | 22    | 3           | 13     | 50         |   |
| 8  | ПК                           | 49                                 | 29    | 9           | 29     | 114        |   |
| 9  | Общий итог                   | 65                                 | 65    | 11          | 49     | 190        |   |
| 10 |                              |                                    |       |             |        |            |   |

Рис. 7.13. Вид сводной таблицы

На рис. 7.13 показана таблица с частотами использования компьютеров разных видов на разных кафедрах. Например, в ячейке E5 на пересечении столбца Социол (SocSci) и Mac (Mac) указано значение 6, т.е. всего шесть из 190 преподавателей кафедры социологии используют на своих учебных курсах компьютеры Macintosh. Какова процентная доля разных компьютеров на разных кафедрах? Попробуем изменить сводную таблицу для отображения этих процентных долей.

Для этого выполните следующие действия:

1. Щелкните правой кнопкой мыши на любой ячейке таблицы, а затем выберите команду Параметры поля (Field Settings) в контекстном меню.
2. В диалоговом окне Вычисление поля сводной таблицы щелкните на кнопке Дополнительно>> для разворачивания нижней части окна.
3. В списке Дополнительные вычисления выберите элемент Доля от суммы по столбцу и щелкните на кнопке ОК. После выполнения этих действий сводная таблица изменится и будет выглядеть так, как на рис. 7.14.

|    | A                            | B         | C       | D          | E       | F          | G |
|----|------------------------------|-----------|---------|------------|---------|------------|---|
| 1  |                              |           |         |            |         |            |   |
| 2  |                              |           |         |            |         |            |   |
| 3  | Количество по полю Компьютер | Кафедра   |         |            |         |            |   |
| 4  | Компьютер                    | Биз. Экон | Матем   | Мед. Здрав | Социол  | Общий итог |   |
| 5  | Мас                          | 6,06%     | 18,46%  | 0,00%      | 12,50%  | 11,58%     |   |
| 6  | Малый компьютер              | 1,52%     | 3,08%   | 0,00%      | 2,88%   | 2,11%      |   |
| 7  | Майнфрейм                    | 18,18%    | 33,85%  | 27,27%     | 27,88%  | 26,32%     |   |
| 8  | ПК                           | 74,24%    | 44,62%  | 72,73%     | 56,33%  | 60,00%     |   |
| 9  | Общий итог                   | 100,00%   | 100,00% | 100,00%    | 100,00% | 100,00%    |   |
| 10 |                              |           |         |            |         |            |   |

Рис. 7.14. *Nāī āī āy ōāāēēōā n ī ōī ōāī ōī ūī ē āī ēyī ē -ānōī ōū ēnī ī ēūcī āāī ēy ōācī ūō ēī ī -ī ūōāōāī ā ī ā ōācī ūō ēāōāāōā*

Процентная доля персональных компьютеров велика для кафедр бизнеса и экономики, медицины и здравоохранения, а процентная доля компьютеров Macintosh относительно велика для кафедры математики. Действительно, на кафедре бизнеса и экономики используется много персональных компьютеров (значение 74,24% в ячейке B8) и мало компьютеров Macintosh (значение 6,06% в ячейке B5). По сравнению с ней на кафедре математики используется чуть меньше персональных компьютеров (значение 44,62% в ячейке D8) и чуть больше компьютеров Macintosh (значение 18,46% в ячейке D5). Нас интересует отношение процентных долей использования компьютеров Macintosh на этих кафедрах. Значение 18,46% (процентная доля компьютеров Macintosh на кафедре математики) меньше, чем 44,62% (процентная доля персональных компьютеров на кафедре математики), но больше, чем 6,06% (процентная доля компьютеров Macintosh на кафедре бизнеса и экономики). Попробуем изменить таблицу таким образом, чтобы процентная доля компьютеров Macintosh вычислялась по отношению к частоте использования персональных компьютеров, а не всех компьютеров. В таком случае будет проще анализировать соотношение персональных компьютеров и компьютеров Macintosh.

*Āēy āū-ēnēāī ēy ī ōī ōāī ōī ī ē āī ēē ā ēāāāī ē ū-āēēā ī ī ī ōī ōāī ēp ē -ānōī ōā ēnī ī ēūcī -āāī ēy ī āōnī ī āēūī ūō ēī ī ī ūōāōāī ā āūī ī ēī ēōā ī āōā-ēnēāī ī ūā ī ēāā āāēnōāēy.*

1. Щелкните правой кнопкой мыши на любой ячейке с указанием процентной доли, а затем выберите команду Параметры поля в контекстном меню.
2. В списке Дополнительные вычисления выберите элемент Доля (% Of).
3. В списке поле (Base field) выберите элемент Компьютер (Computer), если он еще не выбран.
4. В списке элемент (Base item) выберите элемент ПК (PC). После выполнения этих действий диалоговое окно Вычисление поля сводной таблицы будет выглядеть так, как на рис. 7.15.
5. Щелкните на кнопке ОК.

На рис. 7.16 показан окончательный вид сводной таблицы после выполнения перечисленных выше действий с отображением в каждой ячейке процентной доли по отношению к частоте использования персональных компьютеров отдельно для каждой кафедры.

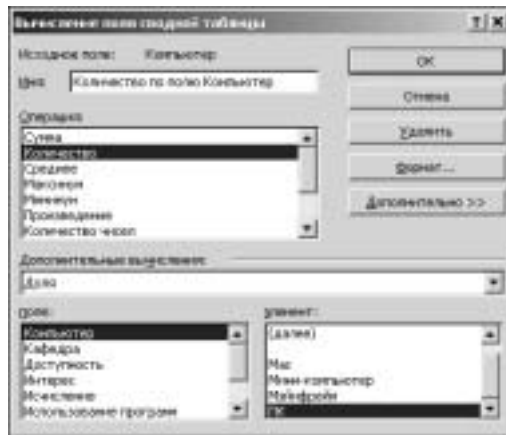


Рис. 7.15. Вычисление поля сводной таблицы

|    | A                            | B         | C       | D          | E       | F          | G |
|----|------------------------------|-----------|---------|------------|---------|------------|---|
| 1  |                              |           |         |            |         |            |   |
| 2  |                              |           |         |            |         |            |   |
| 3  | Количество по полу Компьютер | Кафедра   |         |            |         |            |   |
| 4  | Компьютер                    | Биз. Экон | Матем   | Мед. Здрав | Социол  | Общий итог |   |
| 5  | Mac                          | 8,16%     | 41,38%  | 0,00%      | 21,43%  | 19,30%     |   |
| 6  | Мини-компьютер               | 2,04%     | 6,90%   | 0,00%      | 3,57%   | 3,51%      |   |
| 7  | Майнфрейм                    | 24,49%    | 75,66%  | 37,50%     | 46,43%  | 43,86%     |   |
| 8  | ПК                           | 100,00%   | 100,00% | 100,00%    | 100,00% | 100,00%    |   |
| 9  | Общий итог                   |           |         |            |         |            |   |
| 10 |                              |           |         |            |         |            |   |

Рис. 7.16. Расчет доли компьютеров Macintosh по кафедрам

На кафедре бизнеса и экономики компьютеры Macintosh составляют 8,16% от персональных компьютеров (ячейка B5), на кафедре математики — 41,38% (ячейка D5), а в общем компьютеры Macintosh составляют почти пятую часть от всего количества персональных компьютеров (ячейка F5). Более высокая процентная доля персональных компьютеров на кафедре бизнеса и экономики, по-видимому, связана с их широким распространением в сфере бизнеса и экономики. Аналогично, более высокая процентная доля компьютеров Macintosh на кафедре математики, вероятно, объясняется их популярностью среди математиков.

× 0,0816 = 0,0816 \* 100 = 8,16%. Аналогично для других кафедр.

1. Щелкните правой кнопкой мыши на любой ячейке с указанием процентной доли, а затем выберите команду Параметры поля в контекстном меню.
2. В списке Дополнительные вычисления выберите элемент Нет (Normal) и щелкните на кнопке ОК.

## Вычисление ожидаемого количества наблюдений

Если процентное распределение разных видов компьютеров одинаково внутри каждой кафедры, то процентные доли от суммы по столбцу (см. рис. 7.14) должны быть почти одинаковыми для всех кафедр. В таком случае можно предположить, что ка-

факультета и выбор типа компьютера *не зависят друг от друга*, поэтому распределение разных видов компьютеров не зависит от факультета. С другой стороны, если для разных факультетов существует разное распределение видов компьютеров, то факультет и выбор типа компьютера *зависят друг от друга*. В таком случае выводы об использовании разных видов компьютеров следует делать для каждого факультета отдельно.

Ранее было получено свидетельство того, что процентная доля компьютеров Macintosh отличается для двух факультетов. Для определения статистической значимости этой разницы сформулируем следующие гипотезы:

- нулевая гипотеза  $H_0$ : распределение разных видов компьютеров одинаково внутри каждого факультета;
- альтернативная гипотеза  $H_a$ : распределение разных видов компьютеров зависит от факультета.

Как проверить нулевую гипотезу? Для этого нужно протестировать распределение разных видов компьютеров на разных факультетах и сравнить фактические значения с ожидаемыми, полученными из предположения, что распределение не зависит от факультета.

Как вычислить ожидаемые значения? Согласно нулевой гипотезе распределение разных видов компьютеров одинаково внутри каждого факультета. Наилучшая оценка этого распределения основана на суммарных процентных долях для всех факультетов по отношению к суммарному количеству ответов (т.е. на значениях в столбце F на рис. 7.14). Таким образом, предполагается, что на 11,58% учебных курсов используются компьютеры Macintosh, на 26,32% — мэйнфреймы и т.д. Для выражения этих данных в виде количества наблюдений нужно умножить процентные доли на общее количество учебных курсов на каждом факультете. Например, на факультете математики предлагается 65 учебных курсов, поэтому ожидаемая доля 11,58% компьютеров Macintosh соответствует следующему количеству учебных курсов:  $65 \times 0,1158 = 7,53$ . Обратите внимание, что фактическое значение равно 12 (в ячейке B5 на рис. 7.13), т.е. на факультете математики используется больше компьютеров Macintosh, чем следовало бы ожидать. Данное вычисление эквивалентно следующей формуле:

$$\text{ожидаемое количество} = \frac{(\text{сумма по строке}) \times (\text{сумма по столбцу})}{\text{общее количество наблюдений}}$$

Таким образом, ожидаемое количество наблюдений для компьютеров Macintosh на факультете математики определяется формулой

$$\text{ожидаемое количество} = \frac{22 \times 65}{190} = 7,53.$$

Для создания таблицы с ожидаемыми значениями можно использовать формулы программы Excel или инструменты модуля StatPlus.

Алгоритм действий:

1. Выберите команду меню StatPlus⇒Descriptive Statistics⇒Table Statistics (StatPlus⇒Описательная статистика⇒Статистика таблицы).
2. В появившемся диалоговом окне Analyze a Two Way Table (Анализ табличных данных) укажите диапазон ячеек A4 : E8.
3. Щелкните на кнопке Output и в диалоговом окне Output Options выберите переключатель New Worksheet. В текстовом поле введите строку Распределение компьютеров (Computer Department Table) и щелкните на кнопке ОК.
4. Щелкните на кнопке ОК для создания таблицы с ожидаемыми значениями, которая показана на рис. 7.17.

|    | A                      | B        | C     | D         | E      |
|----|------------------------|----------|-------|-----------|--------|
| 1  | Table Statistics       |          |       |           |        |
| 2  | <b>Observed Counts</b> | Биз Экон | Метел | Аад Эдрел | Соцпол |
| 3  | Мас                    | 4        | 12    | 0         | 6      |
| 4  | Машинкомпьютер         | 1        | 2     | 0         | 1      |
| 5  | Майндрейт              | 12       | 22    | 3         | 13     |
| 6  | ГК                     | 48       | 23    | 8         | 28     |
| 7  |                        |          |       |           |        |
| 8  | <b>Expected Counts</b> | Биз Экон | Метел | Аад Эдрел | Соцпол |
| 9  | Мас                    | 7.84     | 7.53  | 1.27      | 5.56   |
| 10 | Машинкомпьютер         | 1.38     | 1.37  | 0.23      | 1.01   |
| 11 | Майндрейт              | 17.37    | 17.11 | 2.89      | 12.63  |
| 12 | ГК                     | 39.68    | 39.03 | 6.60      | 28.80  |

Dep. 7.17. 0aææöü ñ 0aæðè-àpèèì è è òæèààì ùì è çìá-  
-áí èyì è

Помимо показанных на рис. 7.17 таблиц с фактическими и ожидаемыми значениями, эта команда создает дополнительные данные, которые подробно рассматриваются далее в главе. В данном случае таблица Expected Counts (Ожидаемые значения)



Àëý èñííëùçíááíëý ó-ááííáí ííñíáëý **Распределения.xls (Distributions.xls)** áùííëíëðà í áðá-èñëáíí úà íëæà ááëñoáëý.

1. Откройте учебное пособие **Распределения.xls (Distributions.xls)**, которое находится в каталоге **Учебные пособия (Explore)** прилагаемого компакт-диска. Включите режим поддержки макросов.
2. Выберите раздел **хи-квадрат (Chi-squared)** и ознакомьтесь с материалом раздела, прокручивая его содержимое вверх или вниз (рис. 7.18).

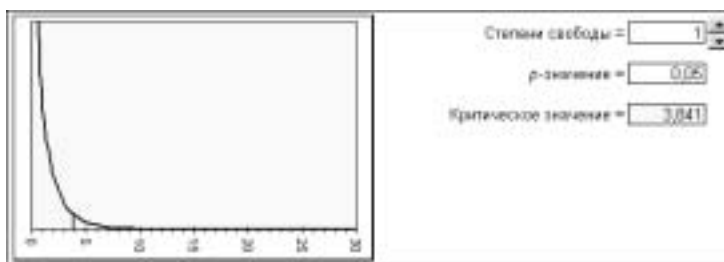


Рис. 7.18. Расчет хи-квадрат

В отличие от нормального и t-распределения,  $\chi^2$ -распределение ограничено значениями  $\geq 0$ . Однако, аналогично t-распределению,  $\chi^2$ -распределение имеет только один параметр — количество степеней свободы. При малом количестве степеней свободы распределение имеет очень асимметричную форму, а при увеличении количества степеней свободы форма распределения становится все более симметричной. Понаблюдайте за изменениями формы распределения для разного количества степеней свободы.

Àëý óááëë-áíëý èí èë-áñoáà òðáí áí áé ñáí áí áù  $\chi^2$ -ðáñí ðáááëáíëý áùííëíëðà óëáçáí íí á íëæà ááëñoáëà.

1. С помощью стрелок справа от поля **Степени свободы (Degrees of freedom)** введите значение 9.

После этого форма распределения примет такой вид, как на рис. 7.19.

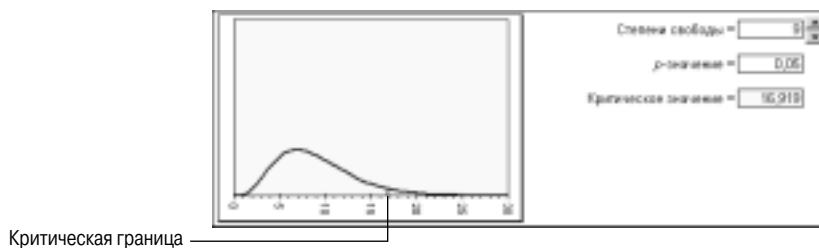


Рис. 7.19. Вид  $\chi^2$ -ðáñí ðáááëáíëý äëý áááýóë òðáí áí áé ñáí áí áù

Аналогично нормальному распределению и t-распределению,  $\chi^2$ -распределение имеет критическую границу для отказа от нулевой гипотезы, но, в отличие от упомя-

На данной диаграмме критическая граница обозначается вертикальной красной линией. В нашем примере она находится у значения  $\alpha=0,05$  и, как показано на рис. 7.19, соответствует значению 16,919. Попробуем изменить значение  $\alpha$  для просмотра его влияния на величину критической границы.

Àëý èçì áí áí èý ááèè÷èí ù èðèèè÷áñèí é áðáí èòù áùíí èí èòá í áðá÷èñèáí í Ùá í èæá ááèñòáèý.

1. В поле p-значение (p-value) введите значение

## Функции программы Excel для работы с $\chi^2$ -распределением

Теперь, зная значение теста и количество степеней свободы, можно проверить нулевую гипотезу. В Excel предусмотрено несколько функций для работы с  $\chi^2$ -распределением. В табл. 7.3 кратко описаны некоторые из них.

Таблица 7.3. Функции Excel для работы с  $\chi^2$ -распределением

| Функция  | Описание   |
|--|--|
| ХИ2РАСП(x; df)<br>(или CHIDIST(x, df) в английской версии)   | Возвращает р-значение для $\chi^2$ -распределения для заданного значения $x$ и степеней свободы $df$   |
| ХИ2ОБР(p; df)<br>(или CHIINV(p, df) в английской версии)   | Возвращает значение $\chi^2$ -распределения для заданного р-значения и степеней свободы $df$   |
| ХИ2ТЕСТ(наблюдаемый_диапазон, ожидаемый_диапазон)<br>(или CHITEST(наблюдаемый_диапазон, ожидаемый_диапазон) в английской версии) | Возвращает значение $\chi^2$ -распределения Пирсона для указанных диапазонов наблюдаемых и ожидаемых значений  |
| PEARSONP(наблюдаемый_диапазон)   | Возвращает значение $\chi^2$ -распределения Пирсона для указанного диапазона наблюдаемых значений. Для его применения требуется подключить модуль StatPlus       |
|  | Возвращает р-значение для $\chi^2$ -распределения Пирсона для указанного диапазона наблюдаемых значений. Для его применения требуется подключить модуль StatPlus |

Полученные ранее данные содержат (помимо прочих параметров) значение  $\chi^2$ -распределения Пирсона, которое равно 14,525, с р-значением, равным 0,105. Поскольку эта вероятность не меньше 0,05, т.е. нет статистически значимого отклонения от гипотезы о независимости, следует принять нулевую гипотезу о том, что распределение видов компьютеров не зависит от кафедр.

## Отклонения от статистики хи-квадрат

Значение статистики хи-квадрат Пирсона основано на каждой ячейке таблицы. Для получения представления о том, какие ячейки вносят наибольший вклад в общее значение, нужно исследовать таблицу стандартизованных разностей.  $\tilde{r}_{ij}$  (*standardized residual*) вычисляется по формуле

$$\text{стандартизованная разность} = \frac{\text{наблюдаемое значение} - \text{ожидаемое значение}}{\sqrt{\text{ожидаемое значение}}}$$

На рис. 7.20 показаны стандартизованные разности для данных об используемых видах компьютеров по кафедрам.

| 14 | Stat Residuals | Экн   | Экон  | Матем | Информ | Систем |
|----|----------------|-------|-------|-------|--------|--------|
| 15 | Mac            | -1,32 | 1,63  | -1,13 | 0,19   |        |
| 16 | Мини-компьютер | -0,33 | 0,54  | -0,48 | -0,01  |        |
| 17 | Мейнфрейм      | -1,28 | 1,18  | 0,06  | 0,10   |        |
| 18 | ПК             | 1,48  | -1,53 | 0,54  | -0,15  |        |

Рис. 7.20. Стандартизованные разности для данных об используемых видах компьютеров по кафедрам

## Другие статистические параметры табличных данных

Очень часто значение  $\chi^2$ -распределения используется для измерения степени взаимосвязи между двумя категориальными переменными. Однако  $\chi^2$ -распределение вместе с р-значением измеряет только *силу* взаимосвязи. Дело в том, что значение  $\chi^2$ -распределения частично зависит от размера выборки и размера таблицы. Например, значение 10  $\chi^2$ -распределения в таблице размером 3×3 является статистически значимым с р-значением, равным 0,04, но значение в таблице размером 4×4 не является статистически значимым с р-значением, равным 0,35.

Степень взаимосвязи переменных в столбце или строке также может измеряться *мерой ассоциации* (measure of association), которая не зависит от размера выборки и размера таблицы. Вообще говоря, чем выше мера ассоциации, тем сильнее взаимосвязь между двумя категориальными переменными.

На рис. 7.21 показаны дополнительные статистические параметры и меры ассоциации, полученные с помощью команды StatPlus⇒Descriptive Statistics⇒Table Statistics (StatPlus⇒Описательная статистика⇒Статистика таблицы) модуля StatPlus.

| 20 | Test Statistics                                   | Value  | df         | p-value |
|----|---|--------|------------|---------|
| 21 | Pearson Chi-Square                                | 14.525 | 9          | 0.105   |
| 22 | Continuity Adjusted Chi-Square                    | 11.806 | 9          | 0.289   |
| 23 | Likelihood Ratio Chi-Square                       | 16.125 | 9          | 0.064   |
| 24 |   |        |            |         |
| 25 | Measures of Association                           | Value  | Std. Error | p-value |
| 26 | Phi   | 0.278  |            |         |
| 27 | Contingency                                       | 0.266  |            |         |
| 28 | Cramer's V  | 0.160  |            |         |
| 29 | Goodman-Kruskal Gamma                             | -0.176 | 0.097      | 0.069   |
| 30 | Kendall's tau-b                                   | -0.111 | 0.062      | 0.073   |
| 31 | Spear's tau-c                                     | -0.092 | 0.052      | 0.073   |
| 32 | Somer's D (CF)                                    | -0.125 | 0.069      | 0.072   |
| 33 | Somer's D (RC)                                    | -0.100 | 0.056      | 0.074   |
| 34 | Warning: More than 1/5 of Fitted Cells are Sparse |        |            |         |

Рис. 7.21. Меры ассоциации и дополнительные статистические параметры

В табл. 7.4 кратко описаны эти параметры и способ их применения.

Таблица 7.4. Статистические параметры таблицы, предусмотренные в модуле StatPlus

| Параметр  | Описание   |
|---|--|
| Pearson Chi-Square ( $\chi^2$ -распределение Пирсона)                           | Вычисляет разницу между наблюдаемым и ожидаемым количеством наблюдений. Приблизительно удовлетворяет $\chi^2$ -распределению с $(r-1) \times (c-1)$ степенями свободы, где $r$ — количество строк в таблице, $c$ — количество столбцов в таблице |
| Continuity Adjusted Chi-Square (Непрерывно подогнанное $\chi^2$ -распределение) | Аналогично $\chi^2$ -распределению Пирсона, за исключением того, что значение $\chi^2$ -распределения подогнано  |
| Likelihood Ratio Chi-Square (Отношение вероятностей $\chi^2$ -распределения)    | Приблизительно удовлетворяет $\chi^2$ -распределению с $(r-1) \times (c-1)$ степенями свободы  |

| Παράμετρος  | Περιγραφή   |
|---|---|
| Phi (Φι-μέτρο)  | Измеряет взаимосвязь между переменными в строке и столбце и находится в диапазоне от -1 до 1. Близкая к 0 величина обозначает отсутствие взаимосвязи  |
| Contingency (Σлучайность)   | Μετρος ассоциации, которая находится в диапазоне от 0 (нет ассоциации) до 1 (максимальная ассоциация). Верхняя граница может быть меньше, чем 1, в зависимости от суммарных значений по столбцам и строкам  |
| Cramer's V (Μετρος ассоциации Крамера)  | Разновидность меры ассоциации, в которой верхняя граница равна 1  |
| Goodman-Kruskal Gamma (Γαμμα-μετρος Γудмана-Κрускала)   | Μετρος ассоциации, которая используется для порядковых переменных в строках и столбцах. Она находится в диапазоне от -1 до 1. Отрицательное значение соответствует отрицательной ассоциации, положительное значение — положительной, а 0 — отсутствию ассоциации  |
| Kendall's tau-b (Ταυ-би метρος Кендалла)  | Μετρος, аналогичная гамма-μετρος Γудмана-Κрускала, за исключением того, что здесь используется коррекция хвостов. Применяется только к порядковым переменным  |
| Stuart's tau-c (Ταυ-σι метρος Стюарта)  | Μετρος, аналогичная таυ-би метρος Кендалла, за исключением того, что здесь используется коррекция размера таблицы. Используется только для порядковых переменных  |
| Somers' D (C R) (Ди-μετρος Сомерса (столбец/строка)) и Somers' D (R C) (Ди-μετρος Сомерса (строка/столбец)) | Разновидность таυ-би меры Кендалла, используемая для порядковых переменных, в которых одна переменная применяется для предсказания значений другой переменной. Ди-μετρος Сомерса (столбец/строка) используется, когда переменная в столбце применяется для предсказания значения другой переменной в строке. И наоборот, Ди-μετρος Сомерса (строка/столбец) используется, когда переменная в строке применяется для предсказания значения другой переменной в столбце |

Ποσκόμυ  $\chi^2$ -распределение является непрерывным и количества наблюдений представляют собой дискретные значения, многие статистики считают, что вместо теста хи-квадрат Пирсона следует применять непрерывно подогнанное  $\chi^2$ -распределение. Авторы полагают, что тест хи-квадрат Пирсона статистически более точен и его можно использовать без подгонки.

Πομίο других статистических параметров, указанных в табл. 7.4, к тесту хи-квадрат Пирсона очень близок тест на основе отношения вероятностей  $\chi^2$ -распределения. Многие статистики предпочитают использовать тест на основе отношения вероятностей  $\chi^2$ -распределения, поскольку он используется в логарифмически-линейном моделировании. К сожалению, подробное описание этой темы выходит за рамки данной книги.

Ни один из трех тестов, показанных на рис. 7.21, не является статистически значимым на уровне 5% (хотя р-значения существенно отличаются). Μετρος ассоциации между распределением видов компьютеров и кафедрами меняется от 0,276 (Φι-μετρος) до 0,160 (μετρος ассоциации Крамера). Ни одна из них не является достаточно большой. Остальные четыре меры ассоциации (γαμμα-μετρος, таυ-би метρος, таυ-σι метρος и ди-μετρος Сомерса) предназначены для порядковых данных и не подходят для номинальных данных.

## Обоснованность применения теста хи-квадрат Пирсона для малых выборок

На рис. 7.21 можно заметить следующее сообщение: Warning: More than 1/5 of Fitted Cells are Sparse (Предупреждение: более 1/5 подогнанных ячеек имеют редкие данные). В этом случае слов *sparse* (редкие) означает, что ячейка содержит значение, величина которого меньше 5. На рис. 7.17 можно заметить, что в таблице с ожидаемыми значениями 6 из 16 ячеек (37,5%) имеют ожидаемые значения менее 5. Данное предупреждение появилось, поскольку их доля 37,5% больше 20% (т.е. 1/5).

Проблема заключается в том, что вряд ли обоснованно применять тест хи-квадрат для таблицы, содержащей большое количество ячеек с очень малыми значениями. Для обоснованного применения теста хи-квадрат Пирсона требуется использовать большие выборки, а потому наличие ячеек с малым количеством наблюдений может создать проблему. Если количество ячеек с ожидаемым значением менее 5 составляет более 20%, то *p*-значение теста хи-квадрат Пирсона может привести к ошибочному принятию или непринятию нулевой гипотезы.

### Исключение редких данных

Что можно сделать для обоснованного применения теста хи-квадрат Пирсона при наличии большого количества редких данных? Для этого можно объединить некоторые строки или столбцы таблицы.

Обратите внимание: в столбце с данными о кафедре медицины и здравоохранения есть три ячейки с редкими данными. Попробуем объединить значения в этом столбце со значениями в другом столбце. С каким именно столбцом следует выполнить это объединение: с данными о кафедре математики или с данными о кафедре социологии? Проще всего было бы исключить данный столбец, но в таком случае об этом исключении обязательно нужно упомянуть в статистическом отчете.

Четыре ячейки с редкими данными находятся в строке о мини-компьютерах, поскольку мини-компьютеры редко используются в учебных курсах по статистике. Однако по своей природе мини-компьютеры очень близки к мэйнфреймам, поэтому объединение данных о них выглядит вполне обоснованным и логичным.

Попробуем создать таблицу с сокращенным количеством ячеек с редкими данными за счет удаления данных о кафедре медицины и здравоохранения и объединения данных о мини-компьютерах и мэйнфреймах. (Кроме того, можно изменить структуру текущей таблицы вместо того, чтобы создавать ее с самого начала, но в таком случае рабочая книга будет содержать две таблицы.)

Ауу пїсааі еу дааееоу п пїедацаі іуі еі ее+апоаі і у+аае п дааееі е ааі іуі е і дап- і ді пдааі а і еє е і і іуі роаді а дасі уо аеа і а і а дасі уо еаоаадао асі і еі еоа і а да+еіа і іуа і еаа ааеіаеу.

1. Выберите команду меню Данные⇒Сводная таблица (Data⇒PivotTable and PivotChart Report).
2. Выберите переключатель в другой сводной таблице или диаграмме (Another PivotTable or PivotChart Report), а затем щелкните на кнопке Далее. Во втором диалоговом окне программы Excel отображается список сводных таблиц из данной книги. Выберите первую сводную таблицу. Щелкните на кнопке Далее.
3. В третьем диалоговом окне мастера сводных таблиц и диаграмм щелкните на кнопке Макет. Перетащите кнопку Компьютер (Computer) в область

Строка (Row), а затем в область Данные (Data); перетащите кнопку Кафедра (Dept) в область Столбец (Column).

- Щелкните на кнопке ОК.
- В третьем диалоговом окне мастера сводных таблиц и диаграмм щелкните на кнопке Параметры.
- Введите значение 0 в текстовом поле для пустых ячеек отображать.
- Щелкните на кнопке Готово.
- Щелкните на стрелке справа возле поля Компьютер (Computer) в сводной таблице и в развернувшемся списке снимите флажок параметра без подписи. Щелкните на кнопке ОК. После этого щелкните на стрелке справа возле поля Кафедра (Dept) в сводной таблице и в развернувшемся списке снимите флажок параметра без подписи. Щелкните на кнопке ОК.
- Щелкните на стрелке справа возле поля Кафедра (Dept) в сводной таблице и в развернувшемся списке снимите флажок Мед\_Здрав (HealthSci) кафедры медицины и здравоохранения.

Теперь 4 из 12 ячеек, т.е. около 33%, имеют редкие данные (содержат величины меньше 5), а 5 из 12 ячеек, т.е. около 42%, содержат величины больше 5. Итак, проблема ячеек с редкими данными еще не решена; для ее решения нужно объединить данные о мини-компьютерах и мэйнфреймах.

## Группирование категорий

Для объединения данных из строк и столбцов следует объединить разные категории в группы. Учтите, что создание группы никак не влияет на исходные данные.

Аёу ётї аёгёдїаагёу ёаоааїдее і ёгё-ётї їүрoаdїа ё їуёг оdаёгїа ауїтёгёоа їаdа-ёёпёагїаї гёаа ааёпoаёу.

- Выделите ячейки A6:A7, т.е. строки с подписями Мини-компьютер (Minicomputer) и Мэйнфрейм (Main).
- Щелкните правой кнопкой мыши на выделении и выберите команду Группа и структура⇒Группировать (Group and Show Detail⇒Group).  
После этого подписи строк в столбце А сместятся в столбец В, а Excel создаст новый столбец Компьютер2 (Computer2) в столбце А в сводной таблице со сгруппированными данными о мини-компьютерах и мэйнфреймах в группе Группа1 (Group1). Создайте для них более информативные имена.
- Выберите ячейку A4, введите строку Группы компьютеров (Computer Groups) и нажмите клавишу <Enter>.
- Выберите ячейку A6, введите строку Мэйнфреймы и мини-компьютеры (Main/Mini), а затем нажмите клавишу <Enter>.

Теперь следует удалить столбец с подписями несгруппированных строк.

- Щелкните на ячейке B4 с подписью Компьютер (Computer), перетащите ее в нижнюю часть таблицы и отпустите кнопку мыши, когда указатель примет форму красного крестика. После этого Excel удалит столбец Компьютер (Computer) из сводной таблицы (рис. 7.22).

Обратите внимание, что только 1 из 9 ячеек (11%) имеет значение меньше 5, а 2 из 9 ячеек (22%) имеют значение 6 или меньше. Реструктуризация таблицы сокращает количество ячеек с редкими данными, но как это отражается на значении вероятности для теста хи-квадрат Пирсона?

В строке Мэйнфреймы и мини-компьютеры (Main/Mini) находится комбинация наблюдений для мэйнфреймов и мини-компьютеров

| Группы компьютеров           | Биз. Экон | Матем | Социал | Общий итог |
|------------------------------|-----------|-------|--------|------------|
| Мэйнфреймы и мини-компьютеры | 5         | 14    | 7      | 26         |
| Мэйнфрейм                    | 12        | 22    | 13     | 47         |
| ПК                           | 49        | 29    | 28     | 106        |
| Общий итог                   | 66        | 65    | 48     | 179        |

Новая группа компьютеров

Рис. 7.22. Настройка таблицы для анализа

Для анализа данных в программе StatPlus необходимо выполнить следующие шаги:

1. Выберите команду меню StatPlus⇒Descriptive Statistics⇒Table Statistics (StatPlus⇒Описательная статистика⇒Статистика таблицы).
2. В появившемся диалоговом окне Analyze a Two Way Table укажите диапазон ячеек A4:D7.
3. Щелкните на кнопке Output и в диалоговом окне Output Options выберите переключатель New Worksheet. В текстовом поле введите имя Распределение компьютеров 2 (Computer Department Table 2) и щелкните на кнопке ОК.
4. Щелкните на кнопке ОК для создания таблицы с ожидаемыми значениями, которая показана на рис. 7.23.

|    | A                              | B         | C         | D       |
|----|--------------------------------|-----------|-----------|---------|
| 1  | Table Statistics               |           |           |         |
| 2  | Observed Counts                | Биз. Экон | Матем     | Социал  |
| 3  | Мэйнфреймы и мини-компьютеры   | 5         | 14        | 7       |
| 4  | Мэйнфрейм                      | 12        | 22        | 13      |
| 5  | ПК                             | 49        | 29        | 28      |
| 6  |                                |           |           |         |
| 7  | Expected Counts                | Биз. Экон | Матем     | Социал  |
| 8  | Мэйнфреймы и мини-компьютеры   | 8.59      | 8.44      | 8.97    |
| 9  | Мэйнфрейм                      | 17.33     | 17.07     | 12.60   |
| 10 | ПК                             | 38.08     | 38.49     | 28.42   |
| 11 |                                |           |           |         |
| 12 | Std Residuals                  | Биз. Экон | Матем     | Социал  |
| 13 | Мэйнфреймы и мини-компьютеры   | -1.48     | 1.48      | 0.01    |
| 14 | Мэйнфрейм                      | -1.28     | 1.19      | 0.11    |
| 15 | ПК                             | 1.59      | -1.53     | -0.08   |
| 16 |                                |           |           |         |
| 17 | Test Statistics                | Value     | df        | p-value |
| 18 | Pearson Chi-Square             | 12.336    | 4         | 0.015   |
| 19 | Continuity Adjusted Chi-Square | 18.353    | 4         | 0.006   |
| 20 | Likelihood Ratio Chi-Square    | 12.628    | 4         | 0.013   |
| 21 |                                |           |           |         |
| 22 | Measures of Association        | Value     | Std Error | p-value |
| 23 | Phi                            | 0.263     |           |         |
| 24 | Contingency                    | 0.254     |           |         |
| 25 | Cramer's V                     | 0.195     |           |         |
| 26 | Goodman-Kruskal Gamma          | -0.230    | 0.102     | 0.023   |
| 27 | Kendalls tau-b                 | -0.143    | 0.065     | 0.027   |
| 28 | Spearman's rho                 | -0.131    | 0.059     | 0.027   |
| 29 | Somer's D (CF)                 | -0.156    | 0.070     | 0.026   |
| 30 | Somer's D (RC)                 | -0.132    | 0.060     | 0.027   |

Рис. 7.23. Результаты анализа

Как видите, р-значение теста хи-квадрат Пирсона существенно изменилось — с 0,105 до 0,015. Поскольку 0,015 меньше, чем 0,05, тест хи-квадрат Пирсона статистически значим на уровне 5%. В таком случае следует отвергнуть нулевую гипотезу о не-



зависимости и принять альтернативную гипотезу о взаимосвязи между кафедрами и видами компьютеров.

Таблица стандартизованных разностей дает подсказку о природе этой взаимосвязи. Обратите внимание, что наиболее низкие значения разностей наблюдаются для учебных курсов на кафедре социологии, т.е. наблюдаемые значения не очень отличаются от ожидаемых, вычисленных на основе нулевой гипотезы. Большие значения разностей в основном наблюдаются для кафедры бизнеса и экономики и кафедры математики. На кафедре бизнеса и экономики персональные компьютеры используются гораздо чаще, чем предсказывает нулевая гипотеза (обратите внимание на положительную разность 1,59). С другой стороны, на кафедре математики чаще, чем предсказывает нулевая гипотеза, используются компьютеры Macintosh (разность равна 1,42), мэйнфреймы и мини-компьютеры (разность равна 1,27).

Это может быть вызвано рядом причин. Одно из объяснений заключается в том, что в момент опроса большая часть программного обеспечения для бизнеса была предназначена для использования на персональных компьютерах, чем на компьютерах Macintosh, мэйнфреймах и мини-компьютерах.

---

## Таблицы с порядковыми переменными

До сих пор рассматривались способы работы с таблицей по двум номинальным переменным. Попробуем теперь рассмотреть способы работы с таблицей по двум порядковым переменным. Для порядковых переменных существуют более мощные тесты, чем тест хи-квадрат Пирсона, который часто не позволяет получить статистически значимые результаты для порядковых переменных.

В качестве примера рассмотрим переменные Исчисление (Calculus) и Зачислено\_В (Enroll\_B). Переменная Исчисление (Calculus) требует знания основ исчисления для данного учебного курса, например она принимает значения Не требуется (Not req) или Требуется (Prereq). В этом аспекте переменная Исчисление (Calculus) является порядковой, хотя ее можно рассматривать и как номинальную. Если переменная принимает только два значения, то ее можно считать порядковой или номинальной, так как эти значения можно расположить в любом порядке. Переменная Зачислено\_В (Enroll\_B) является категориальной и обозначает годовое количество студентов по одной из следующих категорий: 0-50, 51-100, 101-150, 151-200, 201-300, 301-400, 401-500 и 501 и более. Предполагается, что учебные курсы, требующие знания основ исчисления, будут иметь меньшее количество студентов.

## Проверка взаимосвязи между двумя порядковыми переменными

Попробуем проверить взаимосвязь между двумя порядковыми переменными: необходимостью знать основы исчисления и количеством студентов. Для этого сформулируем следующие гипотезы:

- нулевая гипотеза  $H_0$ : распределение студентов не зависит от необходимости знать основы исчисления;
- альтернативная гипотеза  $H_a$ : распределение студентов зависит от необходимости знать основы исчисления.

Для проверки нулевой гипотезы попробуем создать таблицу по двум категориальным переменным — Исчисление (Calculus) и Зачислено\_В (Enroll\_B).

Àëÿ ñî çäáí èÿ òààèèò ïî äàòî ï äðàí áí í Òî àùî î èí èòà ï äðà÷ èñèáí í Òà í èæà äàèòàèÿ.

1. Выберите команду меню Данные⇒Сводная таблица.
2. Выберите переключатель в другой сводной таблице или диаграмме и щелкните на кнопке Далее. Затем в следующем диалоговом окне еще раз щелкните на кнопке Далее.
3. В третьем диалоговом окне мастера сводных таблиц и диаграмм щелкните на кнопке Макет. Перетащите кнопку Зачислено\_В (Enroll\_B) в область Строка (Row), а затем в область Данные (Data); перетащите кнопку Исчисление (Calculus) в область Столбец (Column).
4. Щелкните на кнопке ОК.
5. В третьем диалоговом окне мастера сводных таблиц и диаграмм щелкните на кнопке Параметры.
6. Введите значение 0 в текстовом поле для пустых ячеек отображать и щелкните на кнопке ОК.
7. Убедитесь в том, что выбран переключатель New Worksheet, и щелкните на кнопке Готово.
8. Щелкните на стрелке справа возле полей со списками Исчисление (Calculus) и Зачислено\_В (Enroll\_B), а затем в развернувшемся списке снимите флажок параметра без подписи. В результате сводная таблица будет иметь такой вид, как на рис. 7.24.

| Зачислено_В | Исчисление | Группы | Общий итог |
|-------------|------------|--------|------------|
| 001-050     | 54         | 20     | 74         |
| 051-100     | 62         | 15     | 77         |
| 101-150     | 41         | 6      | 47         |
| 151-200     | 36         | 4      | 40         |
| 201-300     | 26         | 3      | 29         |
| 301-400     | 28         | 1      | 21         |
| 401-500     | 7          | 2      | 9          |
| 501-        | 28         | 6      | 44         |
| Общий итог  | 284        | 57     | 341        |

Рис. 7.24. Сводная таблица с группировкой данных

В созданной таблице Excel автоматически упорядочивает уровни переменной Зачислено\_В (Enroll\_B) по возрастанию. Например, если бы значение 051-100 было записано в виде 51-100, то программа поместила бы его в нижней части таблицы, поскольку цифра 5 идет после 1, 2, 3 и 4.

Обратите внимание: во втором столбце в строках от 151-200 до 401-500 снова возникает проблема редких данных. Однако только 4 из 16 ячеек имеют значения менее 5, поэтому ситуация с редкими значениями в данном случае не очень сложна. Тем не менее для устранения этого недостатка попробуем скомбинировать данные в строках от 151-200 до 401-500.

Для этого выполните следующие действия:

1. Выделите ячейки A9:A11 с подписями категорий от 201-300 до 401-500.
2. Щелкните правой кнопкой мыши на выделении и выберите команду Группа и структура⇒Группировать.
3. Выберите ячейку A4, введите строку Зачислено (Enrollment) и нажмите клавишу <Enter>.
4. Выберите ячейку A9, введите строку 201-500 и нажмите клавишу <Enter>.

- Щелкните на ячейке B4, перетащите ее в нижнюю часть таблицы и отпустите кнопку мыши, когда указатель примет форму красного крестика. После этого Excel удалит данное поле из сводной таблицы (рис. 7.25).

| Зачислено  | Исчислено | Не требуется | Требуется | Общий итог |
|------------|-----------|--------------|-----------|------------|
| 001-050    |           | 54           | 20        | 74         |
| 051-100    |           | 62           | 15        | 77         |
| 101-150    |           | 41           | 6         | 47         |
| 151-200    |           | 38           | 4         | 40         |
| 201-500    |           | 53           | 6         | 59         |
| 501-       |           | 38           | 6         | 44         |
| Общая итог |           | 284          | 57        | 341        |

Рис. 7.25. Объединенные значения

Попробуем теперь вычислить статистические параметры новой таблицы.

Анализ данных в Excel

(Учтите, что действия будут корректно выполнены с региональным стандартом Английский (США). Поэтому для его установки в операционной системе Windows XP следует выбрать команду Пуск⇒Панель управления, затем в диалоговом окне Панель управления нужно выбрать элемент Язык и региональные стандарты, после чего во вкладке Региональные параметры диалогового окна Язык и региональные стандарты выбрать в списке элемент Английский (США). По окончании работы с примером следует с помощью тех же действий вернуться к текущему региональному стандарту. — I dei . daa.)

- Выберите команду меню StatPlus⇒Descriptive Statistics⇒Table Statistics.
- В появившемся диалоговом окне Analyze a Two Way Table укажите диапазон ячеек A4:C10.
- Щелкните на кнопке Output и в диалоговом окне Output Options выберите переключатель New Worksheet. В текстовом поле введите имя Статистика зачисления (Enrollment Statistics) и щелкните на кнопке ОК.
- Щелкните на кнопке ОК.

После этого таблица будет выглядеть так, как на рис. 7.26.

| Test Statistics                | Value  | df         | p-value |
|--------------------------------|--------|------------|---------|
| Pearson Chi-Square             | 18.013 | 5          | 0.075   |
| Continuity Adjusted Chi-Square | 7.818  | 5          | 0.167   |
| Likelihood Ratio Chi-Square    | 9.762  | 5          | 0.082   |
| Measures of Association        | Value  | Std. Error | p-value |
| Phi                            | 0.171  |            |         |
| Contingency                    | 0.169  |            |         |
| Cramer's V                     | 0.171  |            |         |
| Spearman Rank Summ.            | -0.001 | 0.171      | 0.999   |
| Kendall's Rank                 | -0.002 | 0.171      | 0.999   |
| Spear's Rho                    | -0.002 | 0.171      | 0.999   |
| Somer's D (CR)                 | -0.002 | 0.171      | 0.999   |
| Somer's D (RC)                 | -0.002 | 0.171      | 0.999   |

Рис. 7.26. Статистические тесты, в которых не учитывается упорядочение данных

Интересно отметить, что в статистических тестах, в которых не учитывается упорядочение данных (тесте хи-квадрат Пирсона, непрерывно подогнанном тесте хи-квадрат, тесте на основе отношения вероятностей), нулевую гипотезу нельзя отверг-

нуть на уровне значимости 0,05. С другой стороны, в статистических тестах, в которых учитывается упорядочение данных (гамма-тесте Гудмана–Крускала, тау-би тесте Кендалла, тау-си тесте Стюарта и ди-тесте Сомерса), нулевая гипотеза успешно отвергается. Эти результаты иллюстрируют важное правило: всегда следует использовать именно те тесты, которые наилучшим образом подходят для характеристик данных. Если полагаться на тесты, учитывающие упорядочение данных, то нулевую гипотезу следует отвергнуть и принять альтернативную гипотезу о том, что распределение студентов зависит от необходимости знать основы исчисления.

Для более глубокого исследования разностей попробуем внимательно изучить таблицу с ожидаемыми значениями и стандартизованными разностями (рис. 7.27).

| Table Statistics |              |           |
|------------------|--------------|-----------|
| Observed Counts  | Не требуется | Требуется |
| 001-050          | 54           | 20        |
| 051-100          | 62           | 15        |
| 101-150          | 41           | 6         |
| 151-200          | 36           | 4         |
| 201-500          | 53           | 6         |
| 501-             | 38           | 6         |
| Expected Counts  | Не требуется | Требуется |
| 001-050          | 61.83        | 12.37     |
| 051-100          | 64.13        | 12.87     |
| 101-150          | 39.14        | 7.86      |
| 151-200          | 33.31        | 6.69      |
| 201-500          | 49.14        | 9.86      |
| 501-             | 36.65        | 7.25      |
| Std. Residuals   | Не требуется | Требуется |
| 001-050          | -0.57        | 2.17      |
| 051-100          | -0.27        | 0.99      |
| 101-150          | 0.30         | -0.66     |
| 151-200          | 0.47         | -1.04     |
| 201-500          | 0.55         | -1.23     |
| 501-             | 0.22         | -0.50     |

Рис. 7.27. Таблица с наблюдаемыми значениями, ожидаемыми значениями и стандартизованными разностями

Из этой таблицы видно, что нулевая гипотеза дает преуменьшенное количество учебных курсов с обязательным знанием исчисления с количеством студентов в диапазоне 1–50, т.е. дает значение 12,37 при наблюдаемом значении 20. По мере возрастания количества студентов нулевая гипотеза дает преувеличенное количество учебных курсов с обязательным знанием исчисления. Например, если нулевая гипотеза верна, то почти 10 учебных курсов с обязательным знанием исчисления должны насчитывать 201–500 студентов, тогда как их наблюдается всего шесть. Из полученных данных можно сделать вывод, что количество студентов на курсе и необходимость знания исчисления зависят друг от друга, т.е. учебные курсы с необходимостью знать основы исчисления насчитывают меньше студентов.

## Специализированный порядок расположения значений

При работе с порядковыми данными их значения в сводной таблице располагаются в определенном порядке. Значения числовой или символьной переменной легко располагаются в алфавитно-цифровом порядке, но как упорядочить значения переменной, которая не является числовой или символьной? Например, значения переменной *Интерес* (Interest), которая выражает степень интереса к дополнительным статистическим материалам, имеют определенный порядок: наименьший (least), ма-

лый (low), некоторый (some), высокий (high), наибольший (most). Однако он не соответствует алфавитно-цифровому порядку расположения. Для решения данной проблемы можно присвоить значениям переменной Интерес (Interest) числовые величины, например 1 = наименьший, 2 = малый, 3 = некоторый, 4 = высокий и 5 = наибольший. Другой подход заключается в создании специализированного порядка расположения значений.

В Excel можно определить произвольное количество специализированных порядков расположения значений. Кроме того, в Excel предусмотрено несколько предварительно определенных порядков расположения значений. Например, встроенный порядок расположения месяцев года (Январь, Февраль, Март, ... , Декабрь) можно использовать для упорядочения названий месяцев. Попробуем создать специализированный порядок расположения значений для переменной Интерес (Interest).

Аэу пї çааг ёу пї äòèàèèçèðí ààí ííàí ïíðуàèà ðàпí íèí æáí ёу çí à-áí èè ï äðàì áí ííè Ин-терес (Interest) áúí í èí èòà ï äðà-èñèáí í úà í èæà äáèñòàèу.

1. Выберите команду меню Сервис⇒Параметры (Tools⇒Options).
2. Выберите вкладку Списки (Custom Lists).
3. Выберите элемент **НОВЫЙ СПИСОК (NEW LIST)** в списке Списки (Custom lists).
4. Введите строку **наименьший (least)** и нажмите клавишу <Enter>.
5. Введите строку **малый (low)** и нажмите клавишу <Enter>.
6. Введите строку **некоторый (some)** и нажмите клавишу <Enter>.
7. Введите строку **высокий (high)** и нажмите клавишу <Enter>.
8. Введите строку **наибольший (most)** и щелкните на кнопке **Добавить (Add)**.

После выполнения этих действий в списке Специализированные списки (Custom lists) диалогового окна Параметры (Options) появится еще один список (рис. 7.28).

9. Щелкните на кнопке **ОК**.

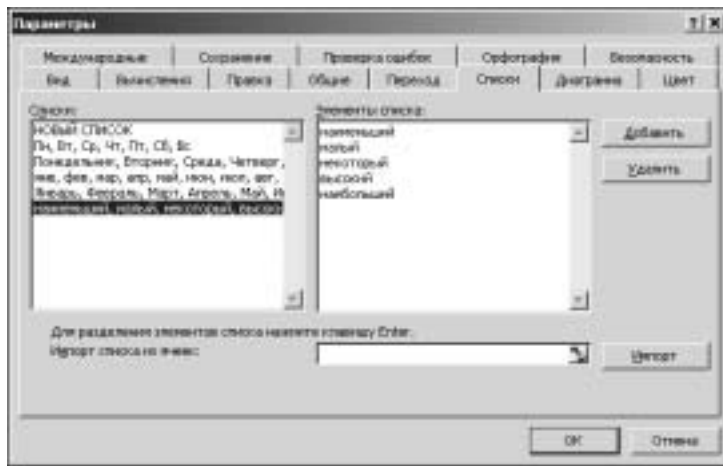


Рис. 7.28. Создание нового списка

Попробуем теперь создать сводную таблицу для значений переменной *Интерес* (Interest) и понаблюдать за автоматическим применением созданного специализированного порядка расположения значений.

×ōīāŪ nī çāāōū nāī āīōīr oāāēēōō āēŷ ī ādāī āīīīē *Интерес* (Interest), āŪīīēīēōā ī ā-ōā-ēñēāī ī Ūā ī ēāā āāēñōāēŷ.

1. Выберите команду меню *Данные*⇒*Сводная таблица*.
2. Выберите переключатель в другой сводной таблице или диаграмме и щелкните на кнопке *Далее*. Затем в следующем диалоговом окне еще раз щелкните на кнопке *Далее*.
3. В третьем диалоговом окне мастера сводных таблиц и диаграмм щелкните на кнопке *Макет*. Перетащите кнопку *Интерес* (Interest) в область *Строка* (Row), а затем в область *Данные* (Data).
4. Щелкните на кнопке *ОК*.
5. В третьем диалоговом окне мастера сводных таблиц убедитесь в том, что выбран переключатель *New Worksheet* и щелкните на кнопке *Готово*.

В результате созданная сводная таблица будет иметь такой вид, как на рис. 7.29.

| Количество поголов <i>Интерес</i> |            |
|-----------------------------------|------------|
| <i>Интерес</i>                    | Итого      |
| наименьший                        | 22         |
| малый                             | 44         |
| некоторый                         | 131        |
| высокий                           | 110        |
| наибольший                        | 71         |
|                                   | 14         |
| <b>Общий итог</b>                 | <b>392</b> |

Рис. 7.29. *Нāī āī āŷ oāāēēōō, ā ēīōīōīē āāōīī āōē-āñēē ī dēī āī ŷāōñŷ nī çāāī ī ŷē oāī āā nī āōōāēēçēōī āāī ī ŷē īīōŷāī ē dānī ī ēī ēāī ēŷ çī ā-āī ēē*

Программа Excel автоматически упорядочивает значения переменной *Интерес* (Interest) не в алфавитно-цифровом, а в заданном прежде специализированном порядке расположения значений: наименьший (least), малый (low), некоторый (some), высокий (high), наибольший (most).

После выполнения этих действий работа с категориальными данными завершается.

Äēŷ çāāāōōāī ēŷ dāāī oŪ ē nīōōāī āī ēŷ īīēō-āī ī ŷō dāçōēīōōāōī ā āŪīīēīēōā ī ādā-ēñēāī ī Ūā ī ēāā āāēñōāēŷ.

1. Сохраните рабочую книгу в каталоге *Примеры* (Student).
2. Выберите команду меню *Файл*⇒*Выход*.

## Упражнения

1. С помощью программы Excel вычислите р-значения перечисленных ниже  $\chi^2$ -распределений.
  - А.  $\chi^2 = 4$ , четыре степени свободы.
  - А.  $\chi^2 = 4$ , одна степень свободы.

- $\bar{A}$ .  $\chi^2 = 10$ , шесть степеней свободы.  
 $\bar{A}$ .  $\chi^2 = 10$ , три степени свободы.
2. С помощью программы Excel вычислите критические значения приведенных ниже  $\chi^2$ -распределений.
- $\bar{A}$ .  $\alpha = 0,10$ , четыре степени свободы.  
 $\bar{A}$ .  $\alpha = 0,05$ , четыре степени свободы.  
 $\bar{A}$ .  $\alpha = 0,05$ , девять степеней свободы.  
 $\bar{A}$ .  $\alpha = 0,01$ , девять степеней свободы.
3. Верно или нет следующее утверждение (и почему): тест хи-квадрат Пирсона измеряет меру ассоциации между одной и другой категориальной переменной?
4. Предположим, что нужно проверить подлинность игральной кости. В табл. 7.5 приведены результаты нескольких подбрасываний кости.

**Таблица 7.5. Результаты подбрасывания кости**

| Сторона | Количество выпадений |
|---------|----------------------|
| 1       | 32                   |
| 2       | 20                   |
| 3       | 28                   |
| 4       | 14                   |
| 5       | 23                   |
| 6       | 15                   |

С помощью команды ХИ2ТЕСТ (CHITEST) проверьте, достаточно ли этих данных, чтобы отвергнуть нулевую гипотезу о подлинности игральной кости.

5. Почему тест хи-квадрат Пирсона не следует использовать для порядковых данных? Какие тесты рекомендуется применять для них?
6. Откройте рабочую книгу Опрос.xls (Survey.xls) и выполните перечисленные ниже задания.
- $\bar{A}$ . Создайте сводную таблицу с указанием кафедр, но без упоминания отсутствующих данных. Напечатайте две версии таблицы: с указанием количества наблюдений и процентных долей.
- $\bar{A}$ . Создайте круговую и столбиковую диаграммы для общего количества наблюдений в сводной таблице Кафедра (Department). На круговой диаграмме укажите наблюдения, а в столбиковой — процентные доли. Напечатайте полученные диаграммы. Какая диаграмма дает более ясное представление о сравнительных размерах групп и почему?
- $\bar{A}$ . Проверьте наличие взаимосвязи между учебными курсами с обязательным знанием исчисления и типом используемых компьютеров с помощью создания сводной таблицы и анализа данных (удалите пустые категории из сводной таблицы). Сформулируйте нулевую и альтернативную гипотезы. Возникает ли в данном случае проблема с редкими данными? Как можно решить эту проблему? Подытожьте полученные результаты и сделайте выводы. Влияет ли условие обязательного знания исчисления на тип используемых компьютеров?

- Ã. Есть ли взаимосвязь между типом используемых компьютеров и количеством студентов на курсе? Создайте сводную таблицу для переменной Зачислено\_В (Enroll\_B) и условия использования персональных или других компьютеров. Объедините данные для всех категорий от 201 до 500 студентов на курсе. Проанализируйте таблицу и сделайте выводы. На основании каких тестов сделаны данные выводы и почему? Является ли количество студентов на курсе определяющим фактором или есть другие более заметные факторы влияния? (İ îãêâçèà. Правда ли, что более сложные учебные курсы насчитывают меньше студентов?)
- Ã. Напечатайте все таблицы и диаграммы. Сохраните рабочую книгу в файле E70прос.xls (E7Survey.xls).
7. В рабочей книге ПреподКолледж.xls (JRCol.xls) собраны данные о заработной плате преподавателей колледжа.
- Ã. Создайте специализированный список должностей в следующем порядке: инструктор (instructor), доцент (assistant professor), адъюнкт-профессор (associate professor) и профессор (full professor).
- Ã. Создайте сводную таблицу со значениями переменной Должность при приеме (Rank Hired) по строкам и значениями переменной Пол (Sex) по столбцам.
- Ã. Проверьте взаимосвязь между должностью и полом. Сформулируйте нулевую и альтернативную гипотезы. Создайте сводную таблицу и изучите статистические параметры таблицы. Какие параметры следует использовать для анализа полученных результатов? Возникает ли проблема редких данных в этом случае? Если да, то как ее можно исправить?
- Ã. Сгруппируйте данные для адъюнкт-профессора (associate professor) и профессора (full professor) и повторите анализ.
- Ã. Сгруппируйте данные для доцента (assistant professor), адъюнкт-профессора (associate professor) и профессора (full professor) и повторите анализ для определения наличия взаимосвязи между полом и парой должностей — инструктор и преподаватель (доцент, адъюнкт-профессор и профессор).
- Ã. Создайте отчет с выводами о полученных результатах и отображением соответствующих таблиц и статистических параметров. Чем отличаются полученные таблицы в трех перечисленных выше случаях? Обсудите проблемы, которые могут возникнуть при попытке исключения редких данных. По вашему мнению, какая из трех таблиц лучше остальных описывает данные? Верно ли, что есть взаимосвязь между должностью и полом преподавателей? Какие фрагменты информации упущены в данном анализе?
- Æ. Создайте еще одну сводную таблицу со значениями переменной Степень (Degree) в разделе данных, значениями переменной Должность при приеме (Rank Hired) по строкам и значениями переменной Пол (Sex) по столбцам. (Поскольку в таблице отображаются наблюдения, то в разделе данных можно использовать значения переменных Должность при приеме (Rank Hired) и Пол (Sex).)
- Ç. С помощью стрелок, расположенных справа от полей, отобразите в таблице данные о преподавателях с ученой степенью Магистр (Master).



- Ё. Вычислите статистические параметры таблицы для данной группы. Связана ли должность в момент приема на работу с полом кандидатов с ученой степенью *Магистр (Master)*. Повторите анализ, если потребуется устранить ячейки с редкими данными.
- Ъ. Создайте отчет с выводами, включая таблицы и статистические параметры. Сохраните полученные результаты в файле *E7ПреподКолледж.xls (E7JRCol.xls)*.
8. Рабочая книга *Простуды.xls (Cold.xls)* содержит данные французского исследования 1961 года для группы из 279 лыжников, которое проводилось в период от 5 до 7 дней. Ежедневно одна группа лыжников получала плацебо (нейтральный солевой раствор), а другая группа — 1 грамм аскорбиновой кислоты. Исследование было предназначено для измерения эффективности действия аскорбиновой кислоты для предотвращения простуды.
- А. Сформулируйте нулевую и альтернативную гипотезы.
- А. Откройте рабочую книгу и проанализируйте данные о простудах.
- А. Подведите итоги и сохраните полученные результаты в файле *E7Простуды.xls (E7Cold.xls)*.
9. Данные в файле *Супруги.xls (Marriage.xls)* содержат информацию о росте новобрачных. В этом исследовании предполагается, что люди стремятся выбирать партнеров такого же роста.
- А. Сформулируйте нулевую и альтернативную гипотезы.
- А. Откройте рабочую книгу и проанализируйте данные в таблице. Какие статистические параметры подходят для анализа этих данных? Следует ли принять или отвергнуть нулевую гипотезу?
- А. Подытожьте полученные результаты и напечатайте статистические параметры в поддержку сделанных выводов. Сохраните полученные результаты в файле *E7Супруги.xls (E7Marriage.xls)*.
10. Откройте рабочую книгу *МненияПол.xls (GenderOP.xls)*, в которой содержатся данные опроса мужчин и женщин по разным социальным и политическим вопросам. Каждый рабочий лист содержит таблицу с ответами на отдельный вопрос.
- А. На каждом листе вычислите статистические параметры таблицы с данными о мнениях.
- А. Для каких вопросов пол респондента связан с результатами ответа?
- А. Подытожьте полученные результаты и объясните, почему на одни вопросы мужчины и женщины дают одинаковые ответы, а на другие — разные.
- А. Сохраните полученные результаты в рабочей книге *E7МненияПол.xls (E7GenderOP.xls)*.
11. Откройте рабочую книгу *МненияРаса.xls (RaceOP.xls)*, в которой содержатся дополнительные данные с ответами представителей разных рас на социальные и политические вопросы.
- А. На каждом рабочем листе вычислите статистические параметры таблицы *Опрос мнений (Opinion Poll)*. Устраните ячейки с редкими данными, комбинируя категории *Черные (Black)* и *Другие (Other)*.
- А. Для каких вопросов полученные результаты не зависят от расы?

- А. Подытожьте полученные результаты и объясните, почему на одни вопросы представители разных рас дают одинаковые ответы, а на другие — разные.
  - А. Сохраните полученные результаты в рабочей книге E7МненияРаса.xls (E7RaceOP.xls).
12. Откройте рабочую книгу Машины.xls (Cars.xls), в которой содержатся данные о разных моделях автомобилей, произведенных в период с 1970 по 1982 год. Попробуйте исследовать зависимость между количеством цилиндров и страной происхождения автомобилей.
- А. Сформулируйте нулевую и альтернативную гипотезы.
  - А. Создайте сводную таблицу со значениями переменной Производитель (Origin) по строкам и значениями переменной Цилиндры (Cylinders) по столбцам.
  - А. Вычислите статистические параметры сводной таблицы. Какие параметры более остальных подходят для анализа данных?
  - А. Есть в таблице ячейки с редкими данными? Устраните их и повторите анализ.
  - А. Подытожьте полученные результаты и сохраните их в рабочей книге E7Машины.xls (E7Cars.xls).
13. Откройте рабочую книгу Дома.xls (HomeData.xls) (см. главу 4, где приводится подробное описание этой рабочей книги).
- А. Проанализируйте данные для определения обоснованности утверждения, что дома в Северо-восточной части пользуются бóльшим спросом, чем дома в остальных районах.
  - А. Подытожьте полученные результаты и сохраните их в рабочей книге E7Дома.xls (E7HomeData.xls).