# Корреляция

# 2

**В ЭТОЙ ГЛАВЕ...**

Как измеряется корреляция

Вычисляем корреляцию

Корреляция и причинно-следственная связь

Ограничение диапазона

Эта книга посвящена регрессионному анализу. Но, прежде чем приступить к рассмотрению регрессии, я намерен обсудить корреляцию. Корреляция — важный компонент регрессионного анализа. Более того, понять суть регрессионного анализа, не зная, что такое корреляция, просто невозможно.

Поэтому, даже если предположить, что вы уже имеете некоторое представление о корреляции, я посвящу несколько страниц обсуждению данной темы. Возможно, вам удастся почерпнуть из изложенного ниже материала кое-что новое для себя.

## Как измеряется корреляция

Часто говорят, что корреляция — это мера зависимости между элементами набора упорядоченных пар. Я не буду особенно строго привязываться к этому определению (несмотря на то, что считаю его содержательным и оно нравится мне). Нам действительно потребуются пары значений, но нет никакой надобности в том, чтобы они были "упорядочены". Наблюдения выступают парами, и коль скоро у вас есть метод, позволяющий устанавливать соответствие между элементами пары, этого будет вполне достаточно.

Важно иметь в виду, что в статистическом анализе корреляцию используют в качестве меры зависимости между переменными, измеряемыми с помощью интервальной шкалы или шкалы отношений. *Интервальной* называют шкалу, в которой числа разделены определенными интервалами. Ее особенностью является то, что нулевую точку можно выбирать произвольно. Классическим примером шкалы этого типа может служить стоградусная температурная шкала: разница температур между 30 и 40 градусами — это то же самое, что

разница температур между 70 и 80 градусами. Одинаковые интервалы представляют одну и ту же разницу значений.

*Шкала отношений* отличается от интервальной шкалы только тем, что она имеет строго определенную нулевую точку. Поскольку на стоградусной шкале нет точки, которая идентифицировалась бы как абсолютный нуль температуры, соответствующий полному прекращению молекулярного движения, она не является шкалой отношений. В то же время шкала Кельвина, имеющая нулевую точку, которой соответствует полное отсутствие тепла, является шкалой отношений, и поэтому мы можем утверждать, что, например, температура 100 градусов Кельвина в два раза больше температуры 50 градусов Кельвина.

В обыденной речи термин "корреляция" в широком смысле означает наличие связи между событиями или явлениями. Однако в контексте регрессионного и других видов статистического анализа под этим подразумевается существование связи между двумя переменными, измеряемыми с помощью интервальной шкалы или шкалы отношений. (Существуют и другие виды корреляционной зависимости, применимые к переменным, измеряемым с помощью *номинальных* или *порядковых* шкал, однако в этом случае используют более строгую терминологию и говорят, например, о *ранговой корреляции* или *двухрядной корреляции*.) В данной книге в целом и в настоящей главе в частности под термином *корреляция*, когда он используется в формальном смысле, будет подразумеваться корреляция, характеризуемая *коэффициентом корреляции Пирсона*.

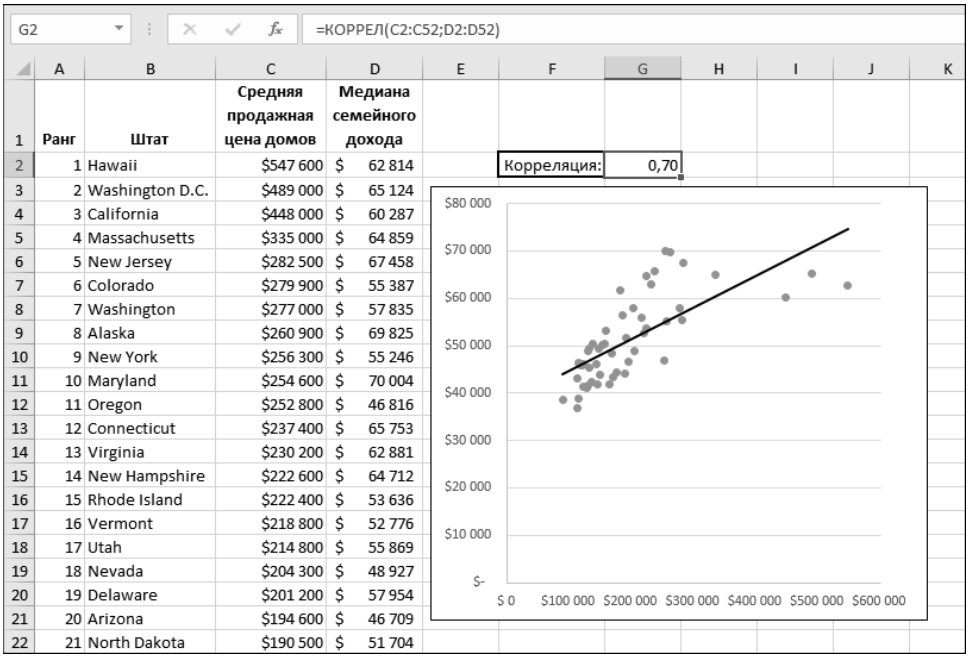
## Выражение степени корреляции

Для выражения степени зависимости случайных величин используют так называемый *коэффициент корреляции*, обозначаемый буквой *r*. Этот коэффициент может принимать значения в интервале от −1,0 до +1,0. Чем ближе *r* к нулю, тем слабее связь. К вопросу о различии между положительными и отрицательными значениями *r* мы вскоре вернемся, а сейчас рассмотрим область его средних значений.

На рис. 2.1 представлены данные о средней продажной цене домов и медиане семейного дохода по каждому из 50 штатов и округу Вашингтон за 2014 год. В ячейке G2 содержится значение коэффициента корреляции между ценой и благосостоянием населения. Как следует из диаграммы, размер дохода и продажная цена изменяются однонаправленно: чем больше медиана семейного дохода в среднем по штату, тем больше средняя продажная цена домов.

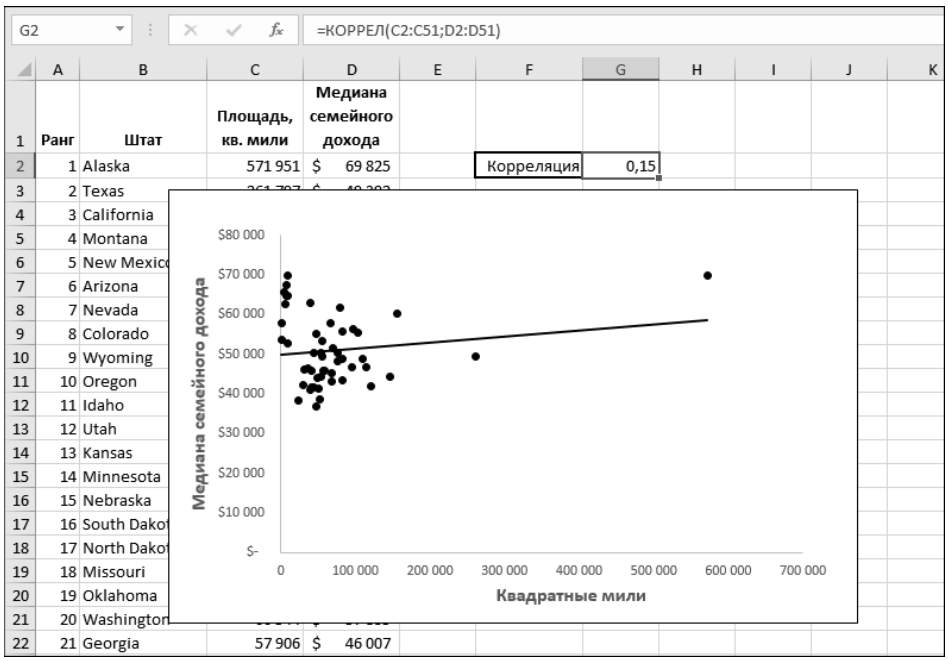
Изображенная на рис. 2.1 прямая диагональная линия называется *линией регрессии*. (В Excel используется альтернативный термин с тем же смыслом — *линия тренда*.) Мы будем неоднократно обращаться к линиям регрессии и их свойствам на протяжении всей книги. А пока что вам достаточно знать, что результаты, получаемые с помощью линейной регрессии, представляются прямой линией и что направление этой линии несет в себе некоторую информацию о природе связи между двумя переменными, а ее положение на диаграмме минимизирует сумму квадратов разностей между точками линии и точками данных диаграммы.

Анализ зависимости между уровнем семейного дохода и стоимостью домов приводит к результатам, которые и можно было ожидать. В штатах, уровень семейных доходов в которых недостаточно высок для того, чтобы цены на жилье могли повышаться, дома стоят дешевле. Сопоставьте между собой результаты, представленные на рис. 2.1 и 2.2.




| | A | B | C | D |
|---|---|---|---|---|
| 1 | Ранг | Штат | Средняя продажная цена домов | Медиана семейного дохода |
| 2 | 1 | Hawaii | $547 600 | $ 62 814 |
| 3 | 2 | Washington D.C. | $489 000 | $ 65 124 |
| 4 | 3 | California | $448 000 | $ 60 287 |
| 5 | 4 | Massachusetts | $335 000 | $ 64 859 |
| 6 | 5 | New Jersey | $282 500 | $ 67 458 |
| 7 | 6 | Colorado | $279 900 | $ 55 387 |
| 8 | 7 | Washington | $277 000 | $ 57 835 |
| 9 | 8 | Alaska | $260 900 | $ 69 825 |
| 10 | 9 | New York | $256 300 | $ 55 246 |
| 11 | 10 | Maryland | $254 600 | $ 70 004 |
| 12 | 11 | Oregon | $252 800 | $ 46 816 |
| 13 | 12 | Connecticut | $237 400 | $ 65 753 |
| 14 | 13 | Virginia | $230 200 | $ 62 881 |
| 15 | 14 | New Hampshire | $222 600 | $ 64 712 |
| 16 | 15 | Rhode Island | $222 400 | $ 53 636 |
| 17 | 16 | Vermont | $218 800 | $ 52 776 |
| 18 | 17 | Utah | $214 800 | $ 55 869 |
| 19 | 18 | Nevada | $204 300 | $ 48 927 |
| 20 | 19 | Delaware | $201 200 | $ 57 954 |
| 21 | 20 | Arizona | $194 600 | $ 46 709 |
| 22 | 21 | North Dakota | $190 500 | $ 51 704 |

***Рис. 2.1.*** *Чем сильнее корреляция, тем ближе к линии регрессии располагаются отдельные наблюдения*




| | A | B | C | D |
|---|---|---|---|---|
| 1 | Ранг | Штат | Площадь, кв. мили | Медиана семейного дохода |
| 2 | 1 | Alaska | 571 951 | $ 69 825 |
| 3 | 2 | Texas | | |
| 4 | 3 | California | | |
| 5 | 4 | Montana | | |
| 6 | 5 | New Mexico | | |
| 7 | 6 | Arizona | | |
| 8 | 7 | Nevada | | |
| 9 | 8 | Colorado | | |
| 10 | 9 | Wyoming | | |
| 11 | 10 | Oregon | | |
| 12 | 11 | Idaho | | |
| 13 | 12 | Utah | | |
| 14 | 13 | Kansas | | |
| 15 | 14 | Minnesota | | |
| 16 | 15 | Nebraska | | |
| 17 | 16 | South Dakot | | |
| 18 | 17 | North Dakot | | |
| 19 | 18 | Missouri | | |
| 20 | 19 | Oklahoma | | |
| 21 | 20 | Washington | | |
| 22 | 21 | Georgia | 57 906 | $ 46 007 |

***Рис. 2.2.*** *Соотношение между медианой семейного дохода по штату и площадью штата носит случайный характер*

На рис. 2.2 показано, как соотносятся между собой две переменные, которые, как ожидается, не связаны между собой. Нет никаких причин предполагать, что между площадью штата и медианой среднего семейного дохода может существовать какая-либо связь. На это же указывает и величина коэффициента корреляции между двумя переменными, равная `0,15` (ячейка G2). Значение `0,15` свидетельствует о слабой корреляции. О слабости этой корреляции говорит и диаграмма, представленная на рис. 2.2. Отдельные точки данных разбросаны случайным образом вокруг линии регрессии. Выброс, соответствующий Аляске, более чем в два раза превышает площадь следующего по величине штата — Техаса, и этот выброс оттягивает правый конец линии регрессии вверх, тем самым придавая коэффициенту корреляции значение, далекое от нулевого, которому соответствует полная независимость случайных переменных.

Зависимость, соответствующая другому концу континуума, настолько сильна, что в природе не встречается (рис. 2.3).

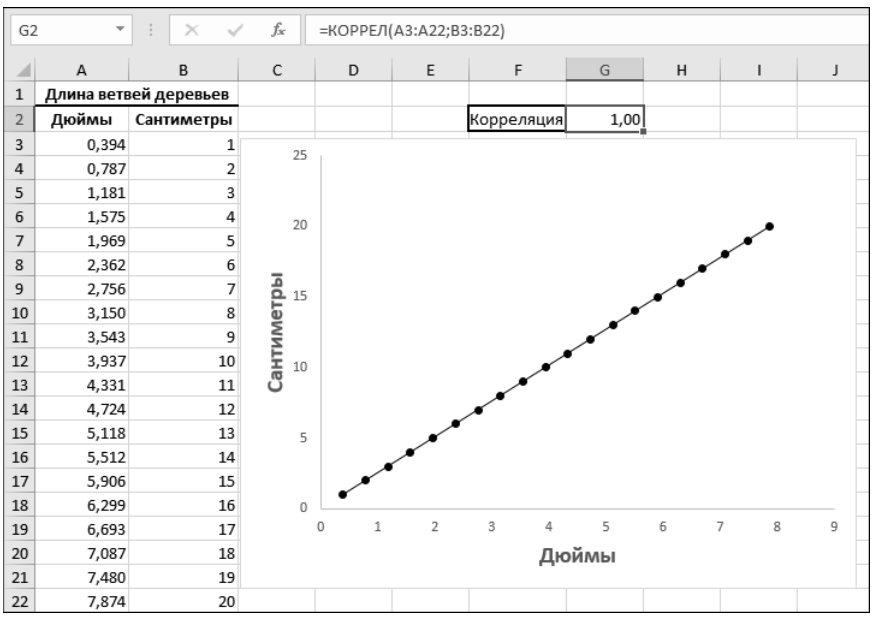



***Рис. 2.3.*** *Эта строгая математическая зависимость совершенно неинформативна, поскольку она настолько сильна, что становится тривиальной*

Зависимость между длиной ветвей деревьев, измеренной в дюймах, и длиной тех же ветвей, измеренной в сантиметрах, характеризуется самой сильной корреляцией, какую только можно получить, с коэффициентом корреляции `1,00` (ячейка G2). Обратите внимание на то, что отдельные точки данных на диаграмме не просто группируются вокруг линии регрессии, а ложатся непосредственно на нее.

Но какой бы сильной ни была эта зависимость, она обусловлена самим способом определения измеряемых величин, а не внутренними характеристиками, общими для обеих переменных (например, длиной ростков, на которую предположительно могут влиять как генетические факторы, так и ранняя подпитка азотными удобрениями).

Поэтому зависимости подобного типа обычно не представляют особого интереса. Однако они позволяют продемонстрировать свойства сильных корреляций, точно так же, как виды взаимосвязи, близкие к случайным, позволяют глубже понять природу слабых зависимостей.

## Определение направления корреляции

Коэффициент корреляции может принимать значения в интервале от −1,0 до +1,0, причем отрицательные значения не только возможны, но и вполне обычны. Знак коэффициента корреляции не имеет никакого отношения к степени зависимости. Он полностью определяется способом измерения переменных.

Рассмотрим пример с бегунами на дистанцию 10 километров. Предположим, вы собрали данные о возрасте бегунов и времени, которое потребовалось каждому из них для преодоления дистанции. Вы обнаружите, что чем меньше времени потребовалось участнику забега для того, чтобы добраться до финиша, тем меньше его возраст. Разумеется, найдется масса примеров противоположного характера. Типичный 10-летний мальчик не пробежит 10000 метров за то же время, что и 18-летний юноша. Однако, вероятнее всего, окажется, что подавляющая часть 60-летних людей не в состоянии сравняться с молодыми 20-летними людьми по скорости бега.

Такая картина соответствует *положительной*, или *однонаправленной*, корреляции. Чем больше значение одной переменной (в данном случае возраста бегуна), тем больше значение другой переменной (количества минут, необходимого для покрытия дистанции от старта до финиша).

А что если бы в вашем распоряжении оказались также данные о количестве часов, еженедельно посвящаемых тренировкам участниками забега на протяжении многих месяцев, предшествующих этому мероприятию? Если бы вы исследовали зависимость между результатами забега на 10 километров и количеством еженедельных тренировочных часов, то, скорее всего, обнаружили бы, что чем больше часов было затрачено на тренировки, тем меньше времени потребовалось бегуну для достижения финишной черты. В данном случае мы имеем дело с *отрицательной*, или *обратной*, зависимостью одной переменной от другой.

Обратите внимание на то, что при анализе обоих видов зависимости — времени пробега дистанции от возраста и времени пробега дистанции от количества тренировочных часов — исследуется соотношение между мерами истекшего времени. Однако эти результаты имеют разный смысл. По достижении определенной точки увеличение возраста уже не сопровождается ускорением бега. С другой стороны, относительное увеличение длительности тренировок обычно сопровождается уменьшением времени пробега дистанции. Результирующая разница между положительной (возраст — время пробега) и отрицательной (количество тренировочных часов — время пробега) корреляцией обусловлена не силой зависимости между двумя переменными, а смыслом и направленностью шкал, используемых для их измерения.

На рис. 2.4 эти зависимости показаны в виде диаграмм.

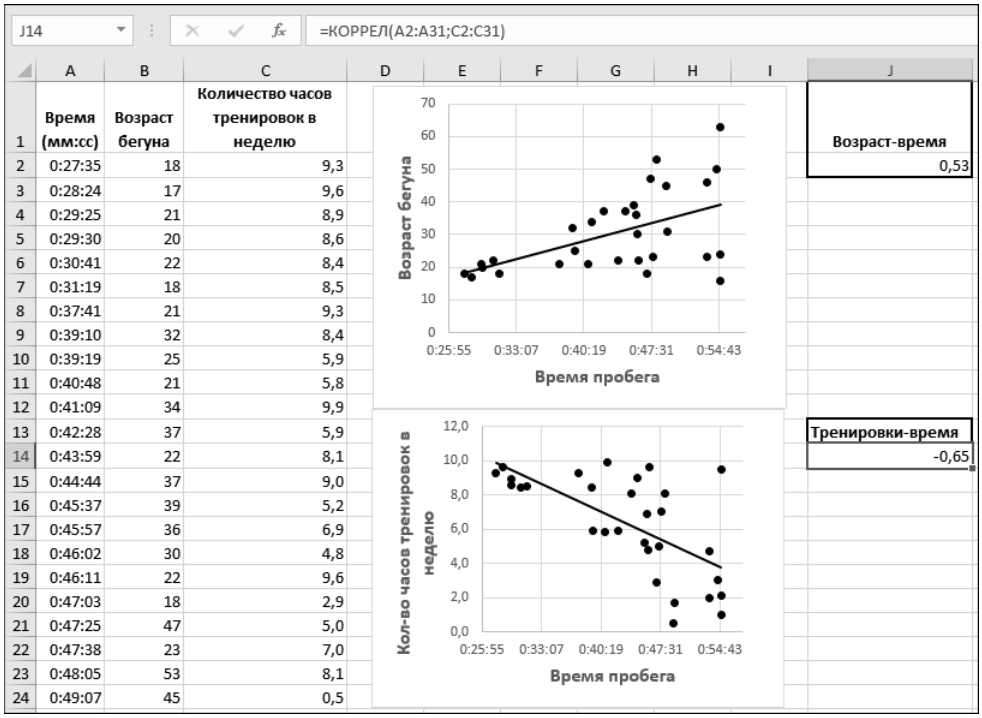



***Рис. 2.4.*** *Направление корреляции находит свое отражение в направлении линии регрессии*

Как следует из рисунка, между возрастом бегуна и временем, в течение которого он преодолевает дистанцию, наблюдается положительная корреляция: как правило, чем старше бегун, тем больше времени ему требуется для того, чтобы добежать до финиша. Значение коэффициента корреляции отображается в ячейке J2. Обратите внимание на наклон линии: она проходит слева направо и снизу вверх.

На рис. 2.4 также видно, что между временем прохождения дистанции и количеством часов еженедельных тренировок наблюдается отрицательная корреляция. В общем случае, чем больше времени отдано тренировкам, тем лучше результат.

Значение коэффициента корреляции отображается в ячейке J14. Обратите внимание на наклон линии регрессии, которая проходит в направлении слева направо и сверху вниз.

Итак, если случайные переменные коррелируют между собой, то коэффициент корреляции может быть как положительным, так и отрицательным.

- В случае положительного коэффициента корреляции линия регрессии проходит на диаграмме в направлении слева направо и снизу вверх. Высокие значения одной переменной сочетаются с высокими значениями другой, а низкие — с низкими.
- В случае отрицательного коэффициента корреляции линия регрессии проходит на диаграмме в направлении слева направо и сверху вниз. Высокие значения одной переменной сочетаются с низкими значениями другой, а низкие — с высокими.

Кстати, порядок, в котором упоминаются переменные, не имеет значения. Скажем, корреляция между ростом и весом — это то же самое, что и корреляция между весом и ростом, причем это касается как степени, так и направления корреляции.

# Вычисляем корреляцию

В главе 1 обсуждалось, чем могут быть полезны такие стандартные оценки, как z-оценка, при описании положения значений в распределениях с использованием определенных стандартных единиц. Узнав от жителя Варшавы, что он зарабатывает 4000 злотых в год, вы вряд ли сочтете эту информацию сколь-нибудь полезной, не будучи знакомым с польской экономикой. Однако если бы вам сказали, что зарплате этого человека в Польше соответствует z-оценка +0,10, то это немедленно указало бы вам на то, что она немного превышает среднюю по стране. (Вспомните о том, что среднее значение z-оценки равно 0,0, а стандартное отклонение — 1,0.)

То же самое справедливо и в отношении коэффициента корреляции. Одним из способов описания зависимости между количеством минут, которое требуется бегуну для преодоления дистанции 10 километров, и возрастом бегуна, о чем говорилось в предыдущем разделе, является использование так называемой *ковариации*. Исследователь, изучающий эти данные, мог бы сообщить вам, что ковариация возраста бегуна со временем пробега 10 километров равна 0,04. Проблема в том, что величина ковариации частично является функцией стандартных отклонений обеих переменных, которые она включает. Поэтому для профессионального аналитика в области легкой атлетики значение ковариации 0,04 еще может нести в себе определенную полезную информацию, точно так же, как информация об уровне зарплаты 4000 злотых может быть содержательной для специалиста в области европейской экономики. Однако ни мне, ни, полагаю, вам эти данные ни о чем не скажут.

Другое дело — коэффициент корреляции. Это стандарт, и на него не влияют различия в измерительных шкалах. Коэффициент корреляции 0,70 описывает положительную зависимость средней степени между двумя переменными, независимо от того, идет ли речь о корреляции между ценой жилья и уровнем семейного дохода или о корреляции между прочностью бумаги на разрыв и процентным содержанием в ней древесины. Это стандартная мера степени и направления числовых зависимостей.

## Первый шаг: ковариация

Несмотря на то что ковариация не столь интуитивно понятна, как корреляция, именно с нее целесообразно начать изучение особенностей корреляционных эффектов. Вот как выглядит одна из формул, позволяющих вычислить ковариацию:

$$s_{xy} = \sum_{i=1}^{N} (X_i - \overline{X})(Y_i - \overline{Y}) / N$$

Возможно, эта формула напоминает вам формулу для дисперсии, которая приводилась в главе 1:

$$S^2 = \sum_{i=1}^{N} (X_i - \overline{X})^2 / N$$

При вычислении дисперсии каждое отклонение от среднего умножается само на себя, т.е. возводится в квадрат. В формуле же для вычисления ковариации отклонение от среднего одной переменной умножается на отклонение от среднего другой переменной. Ковариация случайной величины $X$ с собой равна дисперсии этой величины.

Точно так же, как дисперсия переменной $X$ — это усредненное значение квадратов ее отклонений от среднего значения $X$, ковариация — это усредненное значение произведения отклонений $X$ на отклонения $Y$, где отклонения рассчитываются от соответствующих средних значений.

Посмотрим, как все это работает с конкретными данными. Предположим, вы проводите медицинский эксперимент, цель которого — исследование возможностей снижения медикаментозным путем уровня ЛПНП (липопротеин низкой плотности, так называемый "плохой холестерин") в крови пациентов, страдающих коронарной недостаточностью. В вашем распоряжении оказались данные двух пациентов, Джима и Вирджинии, касающиеся их веса и уровня ЛПНП (рис. 2.5).

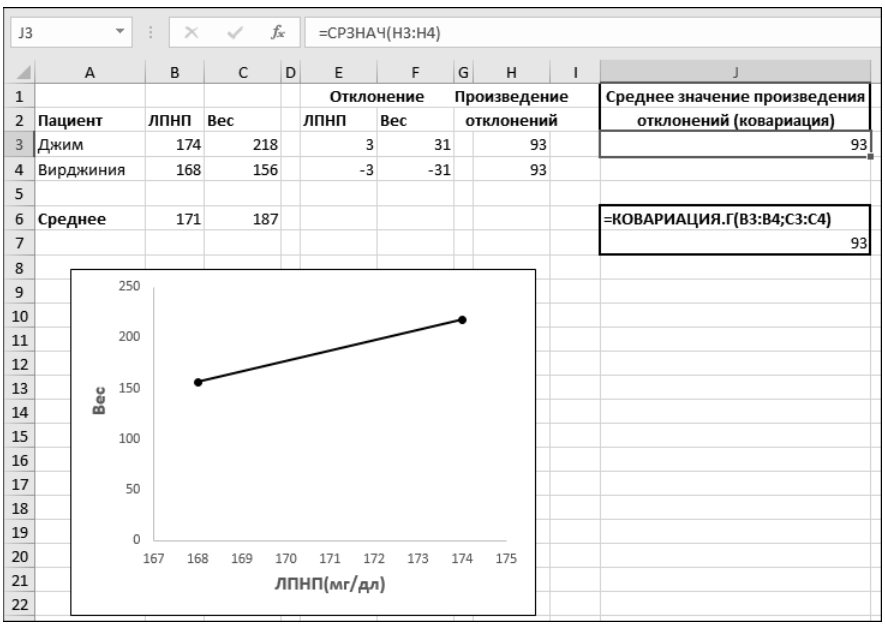



***Рис. 2.5.*** *Вычисление ковариации, а значит, и коэффициента корреляции, требует наличия по крайней мере двух записей и двух случайных переменных*

Результаты измерения уровня ЛПНП и веса Джима отображены в строке 3, Вирджинии — в строке 4.

На рис. 2.5 представлена базовая арифметика расчета ковариации. Исходные необработанные данные содержатся в диапазоне B3:C4. Для расчета отклонений нам нужны средние значения каждой переменной, которые отображаются в ячейках B6 и C6 и вычисляются с помощью функции `СРЗНАЧ()`.

**СОВЕТ**

Планируя корреляционный анализ в Excel, располагайте данные таким образом, чтобы значения разных переменных одной записи находились в одной строке. Например, как показано на рис. 2.5, значения ЛПНП и веса Джима находятся в строке 3. Старайтесь избегать пустых строк. Если предвидится использование инструментов Корреляция или Ковариация для создания матрицы корреляций или ковариаций, избегайте появления пустых столбцов в наборе данных. Во всех других случаях пустые столбцы допустимы (если обратиться к рис. 2.5, то значения переменной `Вес` вполне могли находиться в столбце D, оставляя столбец C пустым), но это может привести к появлению проблем в будущем.

Отклонения получаются вычитанием среднего значения случайной переменной из ее наблюдаемого значения и хранятся в диапазоне E3:F4. Таким образом, значение отклонения ЛПНП Джима, хранящееся в ячейке E3, рассчитывается как разность `B3 – B6`, или `174 – 171`, что дает в результате `3`. Отклонение веса Вирджинии, хранящееся в ячейке F4, рассчитывается как разность `C4 – C6`, или `156 – 187`, что дает в результате `-31`.

Произведения отклонений вычисляются в диапазоне H3:H4 с использованием следующих формул:

- H3 — `=E3*F3`
- H4 — `=E4*F4`

Наконец, ковариация вычисляется путем суммирования произведений отклонений и деления полученного результата на количество наблюдений. Вновь подчеркну, что ковариация — это среднее значение произведений отклонений обеих переменных, аналогично тому, как дисперсия — это среднее значение квадратов отклонений одной переменной.

Для подтверждения правильности расчетов (а также для того, чтобы показать вам, что существует более быстрый способ вычисления ковариации), я использовал функцию `КОВАРИАЦИЯ.Г()` в ячейке J7 для непосредственного получения ковариации. Заметьте, что возвращаемое ею значение совпадает со значением в ячейке J3, рассчитанным более медленным способом.

Подобно тому, как для функций, предназначенных для вычисления дисперсии и стандартного отклонения, в Excel предусмотрены `В`- и `Г`-версии (`ДИСП.В()`, `ДИСП.Г()`, `СТАНДОТКЛОН.В()` и `СТАНДОТКЛОН.Г()`), функция `КОВАРИАЦИЯ()` также имеет две формы, `В` и `Г`, позволяющие вычислять ковариацию соответственно для выборки и генеральной совокупности. Рецепт их применения тот же, что и в случае вычисления дисперсии и стандартного отклонения. Если вы работаете с выборочными данными и хотите оценить параметр генеральной совокупности, используйте `В`-форму:

```
=КОВАРИАЦИЯ.В(M1:M20;N1:N20)
```

В данном случае Excel использует в качестве делителя не *N*, а *(N – 1)*, и поэтому возвращаемое функцией значение не равно в точности среднему значению произведения отклонений. Если же вы работаете с генеральной совокупностью или же не заинтересованы в получении оценки ковариации генеральной совокупности, используйте `Г`-форму:

```
=КОВАРИАЦИЯ.Г(M1:M20;N1:N20)
```

Функция КОВАРИАЦИЯ.Г() использует *N* в качестве делителя и возвращает истинное среднее значение произведения отклонений.

## Учет знаков

Возвращаясь к рис. 2.5, замечу, что он помогает понять, почему при изучении ковариации (а значит, и корреляции) важно учитывать, какой знак она имеет: положительный или отрицательный. Обратимся к четырем оценкам отклонений, содержащимся в диапазоне E3:F4. Эти отклонения всегда вычисляются путем вычитания среднего значения переменной из наблюдаемых значений этой же переменной. Именно поэтому содержащееся в ячейке F3 отклонение 31, полученное вычитанием среднего значения переменной Вес, равного 187, из наблюдаемого значения веса Джима, равного 218, имеет положительный знак. Точно так же, вычитая среднее значение переменной ЛПНП, равное 171, из наблюдаемого значения ЛПНП для Вирджинии, равного 168, мы получаем отрицательную величину -3.

В случае Джима наблюдаемые значения как ЛПНП, так и веса превышают их средние значения, и поэтому оценки обоих отклонений, а вместе с ними и их произведение, имеют положительный знак. В случае же Вирджинии каждое из ее двух наблюдаемых значений меньше среднего значения соответствующей переменной, и поэтому оценки обоих отклонений имеют отрицательный знак, а их произведение — положительный.

Поскольку оба произведения положительны, их среднее также должно быть положительным. Это среднее — ковариация, и, как вы далее увидите, знак ковариации одновременно должен быть и знаком коэффициента корреляции. Как отмечалось ранее, если низкие значения одной переменной ассоциируются с низкими значениями другой (и, как следствие, высокие значения с высокими), мы получаем в результате положительный коэффициент корреляции. В этом случае наклон линии регрессии таков, что она проходит в направлении слева снизу (низкие значения как по горизонтальной, так и по вертикальной оси) и вправо вверх (высокие значения по обеим осям).

Именно такой тип связи переменных наблюдался на рис. 2.5. Рис. 2.6 соответствует противоположной картине.

На рис. 2.6 показатель ЛПНП Джима уменьшен со 174 до 154. В результате этого его отклонение стало отрицательным. А поскольку это изменение приводит также к уменьшению среднего значения ЛПНП в ячейке B6, отклонение показателя ЛПНП Вирджинии становится положительным, так как теперь наблюдаемое значение этой переменной превышает измененное среднее значение.

Конечным эффектом этого единственного изменения показателя ЛПНП Джима является то, что для каждого пациента имеется одна отрицательная оценка отклонения и одна положительная. В этом случае произведения отклонений должны быть отрицательными для обоих пациентов (ячейки H3 и H4 на рис. 2.6). Если все произведения отклонений имеют отрицательные значения, то их среднее — ковариация, а значит, и коэффициент корреляции — также должно быть отрицательным.

Если высоким значениям одной переменной соответствуют низкие значения другой, мы получаем отрицательную корреляцию. В этом случае наклон линии регрессии таков, что она проходит в направлении слева сверху (низкие значения

по горизонтальной оси, высокие — по вертикальной) и вправо вниз (высокие значения по горизонтальной оси, низкие — по вертикальной).

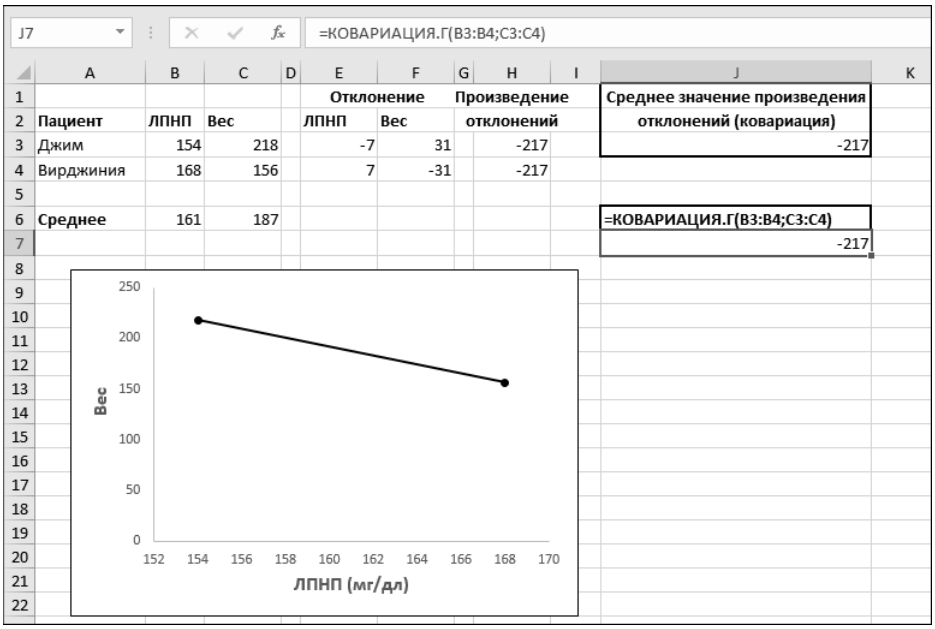



***Рис. 2.6.*** *В результате изменения одного значения ковариация становится отрицательной*

## От ковариации к коэффициенту корреляции

Ковариация всегда имеет тот же знак, что и корреляция, поэтому все, что мы пока знаем, — это то, что в случае двух записей, представленных на рис. 2.5, корреляция между уровнем ЛПНП и весом пациента положительная (тогда как на рис. 2.6 — отрицательная). Смысл *величины* корреляции по-прежнему остается неясным. Тот факт, что ковариация равна 93, мало о чем говорит нам, и, в частности, не ясно, означает это слабую или сильную корреляцию.

Ниже приведена формула, позволяющая вычислить коэффициент корреляции, если вариация известна:

$$r = S_{xy} / (S_x S_y)$$

где

- $r$ — коэффициент корреляции;
- $S_{xy}$ — ковариация;
- $S_x$ и $S_y$ — стандартные отклонения переменных $X$ и $Y$ соответственно.

В этой формуле ковариация делится на стандартные отклонения переменных $X$ и $Y$, тем самым удаляя из ковариации эффекты масштабирования, связанные с единицами измерения.

Стандартное отклонение само создается на основе суммы *квадратов* отклонений и, следовательно, не может быть отрицательным числом. Поэтому, какой бы ни был знак ковариации, он переносится на коэффициент корреляции. Интервал возможных значений *r* включает как положительные, так и отрицательные значения, но ограничивается диапазоном от −1,0 до +1,0 путем исключения размерностей переменных *X* и *Y*, благодаря чему коэффициент корреляции *r* становится безразмерным.

Процесс вычислений продемонстрирован на рис. 2.7.

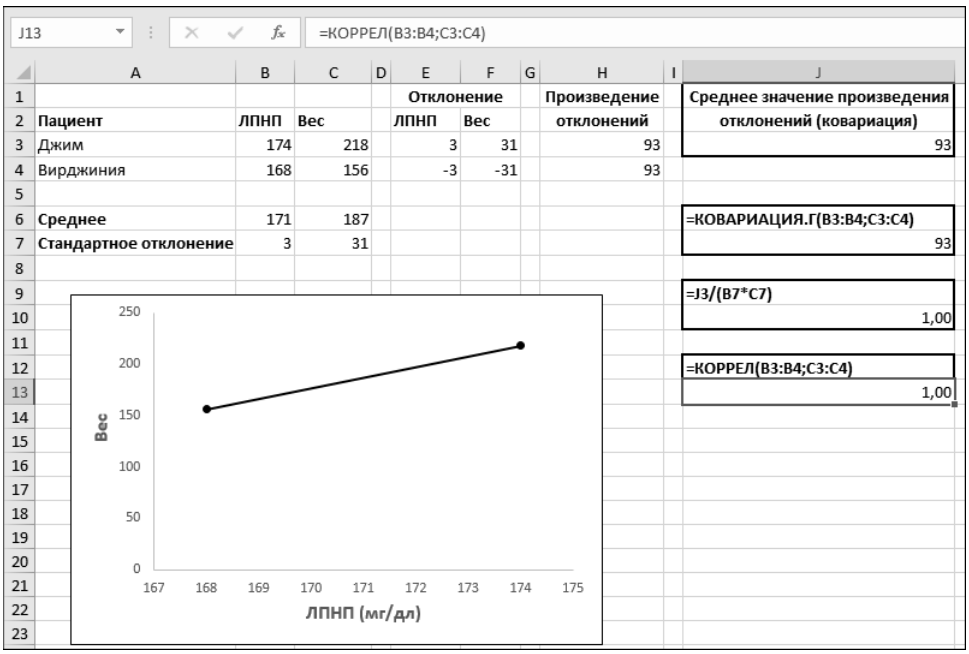
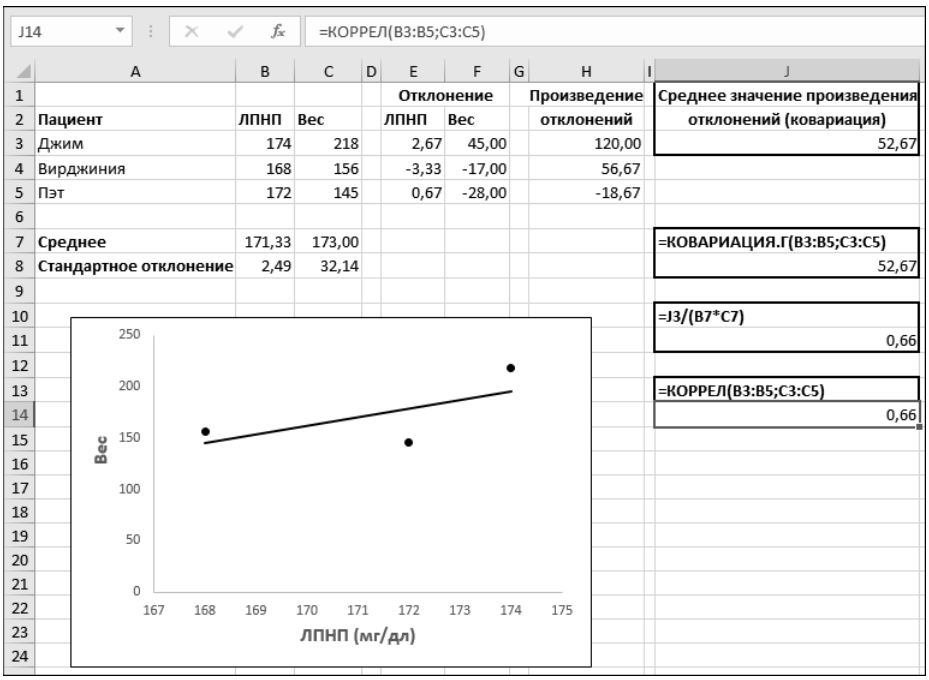



***Рис. 2.7.*** *Можно опустить все промежуточные вычисления, используя функцию* `КОРРЕЛ()`

В ячейке J10 ковариация делится на произведение стандартных отклонений. Соответствующая формула в текстовом виде приведена в ячейке J9. Функция `КОРРЕЛ()` позволяет получить тот же результат с меньшими усилиями, но поэтапное выполнение операций — это единственный способ показать, что происходит внутри "черного ящика".

До сих пор мы рассматривали лишь ситуации, в которых коэффициент корреляции равен либо −1,0, либо +1,0. При наличии только двух записей — в данном случае записей с данными Джима и Вирджинии — корреляция должна быть идеальной. Это можно легко понять, если принять во внимание, что через две точки данных на диаграмме можно провести только одну прямую линию регрессии. Поэтому обе точки должны лежать непосредственно на линии. В этом случае мы имеем дело с идеальной (функциональной) корреляцией. И лишь тогда, когда по крайней мере одна из откладываемых на диаграмме точек находится вне линии регрессии, коэффициент корреляции может быть больше −1,0 и меньше +1,0.

Обратимся к рассмотрению более распространенной ситуации, когда связь между данными не является идеальной. Несмотря на то что организовать наборы данных с результирующим коэффициентом корреляции +1,0 или −1,0 вовсе не трудно, в реальной практике исследований в области медицины, сельского хозяйства, экономики и других областях они не встречаются. Как правило, для коэффициента корреляции характерны значения типа 0,36, 0,53 или 0,71. Чтобы выйти за рамки идеальной корреляции, включим в набор данные третьего пациента в дополнение к предыдущим (рис. 2.8).




| | A | B | C | E | F | H | J |
|---|---|---|---|---|---|---|---|
| 1 | | | | Отклонение | | Произведение | Среднее значение произведения |
| 2 | Пациент | ЛПНП | Вес | ЛПНП | Вес | отклонений | отклонений (ковариация) |
| 3 | Джим | 174 | 218 | 2,67 | 45,00 | 120,00 | 52,67 |
| 4 | Вирджиния | 168 | 156 | -3,33 | -17,00 | 56,67 | |
| 5 | Пэт | 172 | 145 | 0,67 | -28,00 | -18,67 | |
| 6 | | | | | | | |
| 7 | Среднее | 171,33 | 173,00 | | | | =КОВАРИАЦИЯ.Г(B3:B5;C3:C5) |
| 8 | Стандартное отклонение | 2,49 | 32,14 | | | | 52,67 |
| 9 | | | | | | | |
| 10 | | | | | | | =J3/(B7*C7) |
| 11 | | | | | | | 0,66 |
| 12 | | | | | | | |
| 13 | | | | | | | =КОРРЕЛ(B3:B5;C3:C5) |
| 14 | | | | | | | 0,66 |

***Рис. 2.8.*** *При наличии трех и более пар наблюдений корреляция уже не может быть всегда идеальной*

Как следует из рис. 2.8, значение показателя ЛПНП (`172`) Пэт, третьей пациентки, незначительно превышает новое среднее значение (`171,33`), и поэтому оценка отклонения ЛПНП Пэт положительна. Вес Пэт (`145`) меньше среднего значения веса, равного `173`, и поэтому оценка отклонения этого показателя отрицательна. Умножение положительного отклонения на отрицательное дает отрицательное число (`-18,67`). В результате отрицательного вклада этого произведения в общий результат значение коэффициента корреляции снижается с `1,0` (см. рис. 2.7) до `0,66`.

А что если бы показатель ЛПНП Пэт был немного меньше среднего значения ЛПНП? Тогда наряду с отрицательным отклонением веса Пэт от среднего значения отрицательным было бы и отклонение ее показателя ЛПНП, в результате чего произведение обоих отклонений было бы положительным. В этом случае произведение всех трех отклонений также было бы положительным (рис. 2.9).

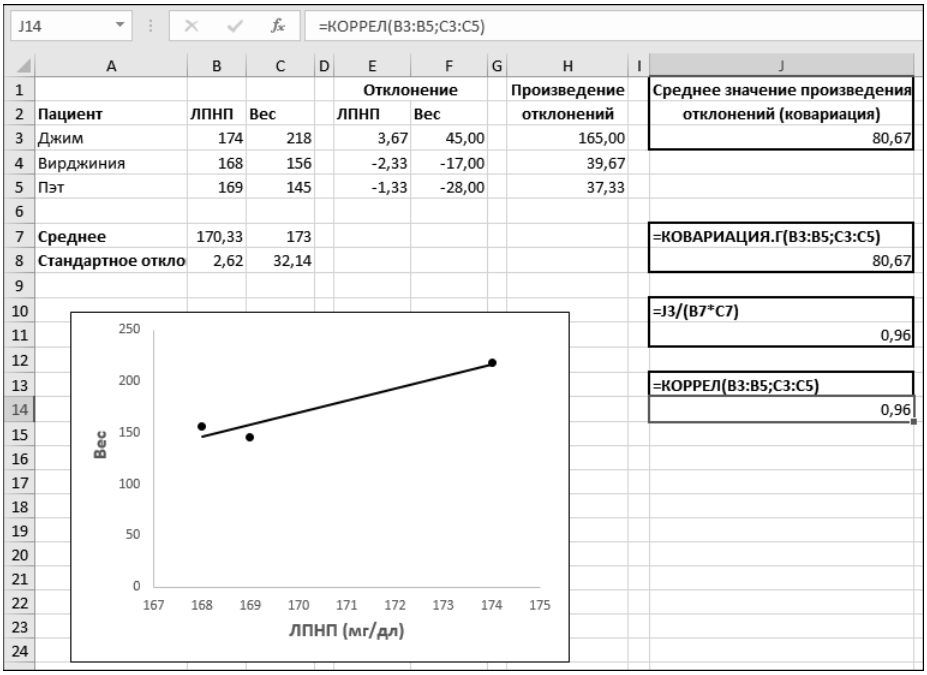



***Рис. 2.9.*** *Все три точки данных на диаграмме находятся вне линии регрессии*

Положительный (отрицательный) знак всех трех произведений отклонений не является свидетельством идеальной корреляции, при которой коэффициент корреляции равен +1,0 или −1,0. Корреляция будет идеальной лишь в том случае, если все точки данных лежат строго на линии регрессии. Мы будем напоминать вам об этом на протяжении всей оставшейся части книги, поскольку от этого напрямую зависит точность прогнозов, основанных на использовании регрессионного анализа.

Тем не менее теперь корреляция очень сильная, на что указывает величина коэффициента `0,96`. Если произведения отклонений положительны (отрицательны) для всех наблюдений, то это указывает на то, что вы получите значение коэффициента, соответствующее сильной корреляции.

## Использование функции КОРРЕЛ()

В предыдущем разделе уже упоминалось о том, что функция `КОРРЕЛ()` может сделать всю работу вместо вас — она сама позаботится о вычислении средних значений, стандартных отклонений и ковариаций и выполнит все необходимые операции умножения, сложения и деления. Например, обратите внимание на то, что результат, отображаемый в ячейке J14 на рис. 2.9, совпадает со значением коэффициента корреляции в ячейке J11, которое является конечным результатом выполнения всех арифметических вычислений.

**ПРИМЕЧАНИЕ**

В Excel имеется другая функция рабочего листа, `ПИРСОН()`, которая принимает те же аргументы и возвращает те же результаты, что и функция `КОРРЕЛ()`. Разумеется, это вносит некоторое неудобство, но функция `ПИРСОН()` используется по крайней мере с середины 1990-х годов, и ради обеспечения обратной совместимости она была оставлена в Excel. Функция названа в честь Карла Пирсона, известного математика, статистика и философа XIX и XX веков, внесшего огромный вклад в развитие прикладной математики вообще и корреляционного анализа в частности. Вы можете использовать любую из этих двух функций, однако в книге я отдаю предпочтение функции `КОРРЕЛ()`, поскольку мне проще вводить ее название.

## О смещении в корреляции

Заметьте, что в названии функции `КОРРЕЛ()` отсутствуют уточняющие элементы `Г` и `В`, которые используются в названиях таких функций, как `СТАНДОТКЛОН()`, `ДИСП()` или `КОВАРИАЦИЯ()`. Например, вызов функции `ДИСП.В()` сообщает Excel о том, что сумму квадратов отклонений следует делить не на количество наблюдений *N*, а на количество степеней свободы, равное $N - 1$. Это позволяет устранить отрицательное смещение выборочной дисперсии и получить несмещенную оценку дисперсии генеральной совокупности, из которой извлечена данная выборка. Например, оценка генеральной дисперсии, получаемая с помощью приведенной ниже формулы, является заниженной и, следовательно, смещенной:

$$S^2 = \sum_{i=1}^{N} (X_i - \bar{X})^2 / N$$

В то же время следующая формула возвращает несмещенное значение генеральной дисперсии:

$$S^2 = \sum_{i=1}^{N} (X_i - \bar{X})^2 / (N - 1)$$

А что можно сказать о выборочном коэффициенте корреляции? Является ли предоставляемая им оценка корреляции смещенной? Да, это так, однако причина смещения иная, нежели в случае дисперсии или стандартного отклонения. Смещение значений, предоставляемых этими статистиками, когда делителем является *N*, обусловлено тем, что сумма квадратов отклонений индивидуальных значений имеет наименьшую величину, если отклонения рассчитываются относительно выборочного среднего, а не любого другого числа, включая среднее по генеральной совокупности. И в силу того, что сумма квадратов отклонений оказывается меньшей, чем следовало бы, выборочная дисперсия дает заниженную оценку дисперсии генеральной совокупности. В итоге выходит так, что деление суммы квадратов отклонений на количество степеней свободы, а не на количество фактических наблюдений, устраняет смещение из выборочной дисперсии.

В отличие от этого смещение коэффициента корреляции, вычисленного на основе выборки, обусловлено скосом выборочного распределения, приводящим к появлению

асимметрии. Если генеральный коэффициент корреляции (часто обозначаемый греческой буквой $\rho$, которая соответствует латинской букве $r$) положителен, асимметричный хвост выборочного распределения $r$ находится слева. Если же $\rho$ имеет отрицательное значение, асимметричный хвост располагается справа. Степень скоса распределения в значительной мере является функцией размера выборки. В случае выборок небольшого размера (скажем, по 10 наблюдений каждая) искажение может быть значительным, однако в случае крупных выборок, включающих тысячи наблюдений, распределение, как правило, выглядит почти симметричным.

В качестве примера на рис. 2.10–2.12 показаны распределения значений выборочных коэффициентов корреляции $r$, полученные на основе выборок по 10 наблюдений каждая, которые извлекались из генеральных совокупностей с истинными коэффициентами корреляции $\rho$, равными соответственно `0,68`, `0,05` и `-0,62`.

На рис. 2.10 значение $\rho$ для генеральной совокупности равно `0,68`. Как видно на диаграмме, выборочные значения $r$ для большинства выборок из этой генеральной совокупности будут находиться в пределах от `0,6` до `0,8`. Однако ниже значения `0,6` еще имеется длинная хвостовая часть распределения, простирающаяся до значения `-0,45`, которое отстоит от значения параметра генеральной совокупности на 1,13 единицы. Но в силу способа вычисления коэффициента корреляции он ограничен справа значением `1,0`, и поэтому на диаграмме выборки со значениями $r$, превышающими `0,68`, плотно группируются в правой хвостовой части распределения.

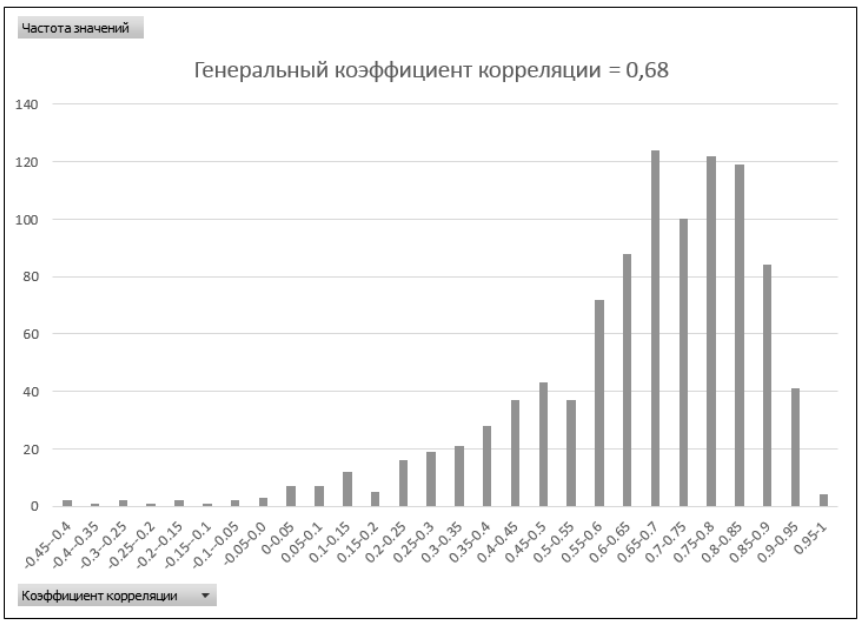



***Рис. 2.10.*** *Распределение выборок с различными значениями коэффициента корреляции при $\rho = 0,68$*

Если генеральный коэффициент корреляции близок к значению `0,0` (рис. 2.11), то по обе стороны от него остается еще много места для распределения выборочных значений коэффициента корреляции, достаточно далеко отстоящих от нуля, и форма распределения приближается к симметричной.

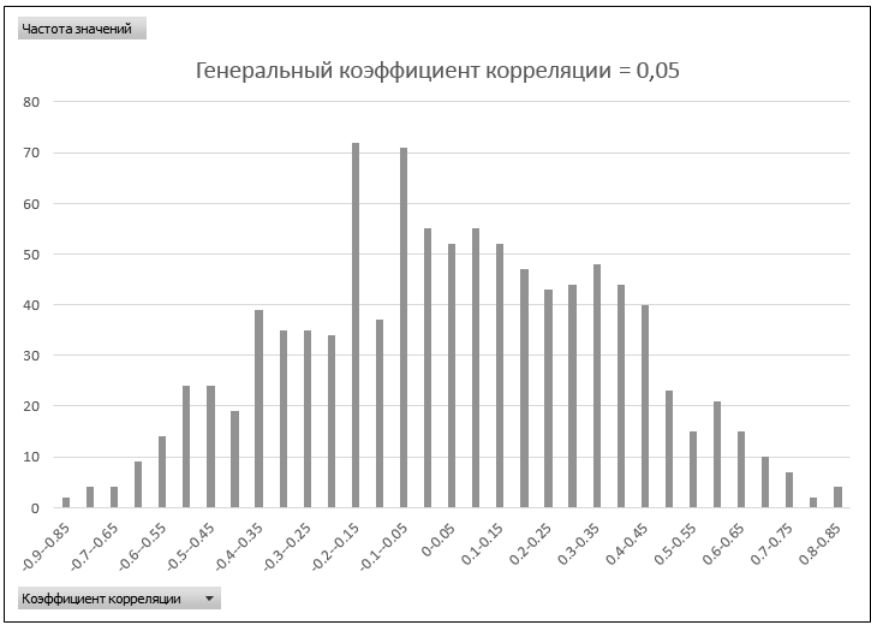



***Рис. 2.11.*** *Распределение выборок с различными значениями коэффициента корреляции при $\rho = 0,05$*

На рис. 2.12 показано, что при отрицательном значении генерального коэффициента корреляции распределение выборочных значений является скошенным, как и на рис. 2.10, но теперь его длинная хвостовая часть находится справа, а не слева.

В случае выборок большего размера эффекты, представленные на рис. 2.10–2.12, были бы менее выраженными. Степень асимметрии распределения была бы намного меньшей, а диапазон значений выборочного коэффициента корреляции — намного более узким. На этом основывается коррекция коэффициента *множественной* корреляции, о чем пойдет речь в главе 5.

Итак, в зависимости от величины $\rho$, коэффициент корреляции $r$ дает смещенную оценку, причем эффект смещения усиливается с уменьшением размера выборок. Тем не менее мы не пытаемся корректировать это смещение аналогично тому, как, например, делали это при вычислении стандартного отклонения, когда подставляли в соответствующую формулу не количество наблюдений, а количество степеней свободы. В действительности количество наблюдений, используемое для вычисления ковариации, не оказывает никакого влияния на величину. Формула

$$r = S_{xy} / (S_x S_y)$$

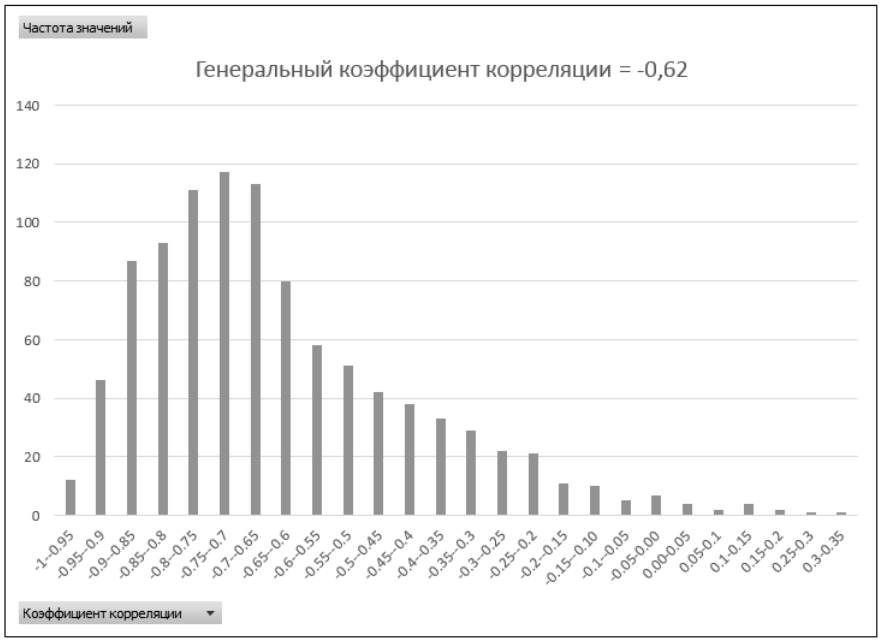



***Рис. 2.12.*** *Распределение выборок с различными значениями коэффициента корреляции при ρ = –0,62*

(уже приводившаяся в этой главе в качестве метода расчета *r*) полезна в качестве метода концептуализации связи между различными способами описания наборов данных (и, по моему мнению, обеспечивает удобный способ вычисления коэффициента корреляции в тех случаях, когда величины ковариации и стандартных отклонений известны). Но для того чтобы увидеть, почему количество наблюдений не влияет на конечный результат вычислений, нужно переписать эту формулу в следующем виде:

$$r = (\sum x_i y_i \,/\, N) \,/\, \sqrt{\sum x_i^2 \,/\, N} \sqrt{\sum y_i^2 \,/\, N}$$

Здесь *x* и *y* представляют отклонения наблюдаемых значений *X* и *Y* от соответствующих средних значений. Переместим *N* из выражения для ковариации в знаменатель формулы:

$$r = (\sum x_i y_i) \,/\, (N \sqrt{\sum x_i^2 \,/\, N} \sqrt{\sum y_i^2 \,/\, N})$$

Произведение квадратных корней из *N*, входящих в выражения для стандартных отклонений в знаменателе, взаимно сокращается с *N*, перешедшим из выражения для ковариации:

$$r = \sum x_i y_i \,/ \sqrt{\sum x_i^2} \sqrt{\sum y_i^2}$$

Иными словами, коэффициент корреляции можно рассчитать как отношение суммы произведений отклонений *X* и *Y* к произведению корня квадратного из суммы квадратов отклонений *X* и корня квадратного из суммы квадратов отклонений *Y*. (Сейчас у вас была отличная возможность убедиться в том, насколько полезным может быть переход от словесных формулировок к формулам.) Обратите внимание на то, что *N* не входит в конечную формулу для *r*. Вы могли использовать *(N − 1)* вместо *N* в промежуточных вычислениях, но это никак не сказалось бы на конечном результате.

## Проверка линейности и наличия выбросов в корреляции

Стандартный коэффициент корреляции — вариант, разработанный Карлом Пирсоном примерно в 1900 году, который обычно и имеют в виду, когда говорят о корреляции в контексте статистики, — предназначен для использования с переменными, связанными между собой линейным соотношением. Хороший пример линейного соотношения представлен на рис. 2.1, из которого следует, что между предлагаемой ценой жилья и уровнем семейного дохода существует линейная зависимость средней степени.

Пример типа зависимости между переменными, для описания которого коэффициент корреляции Пирсона не предназначен, показан на рис. 2.13.

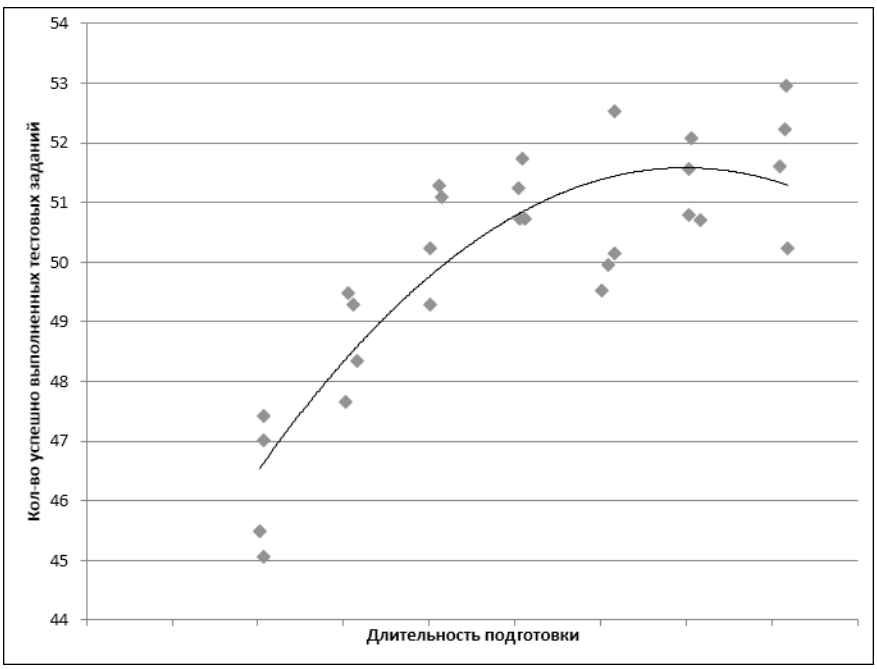


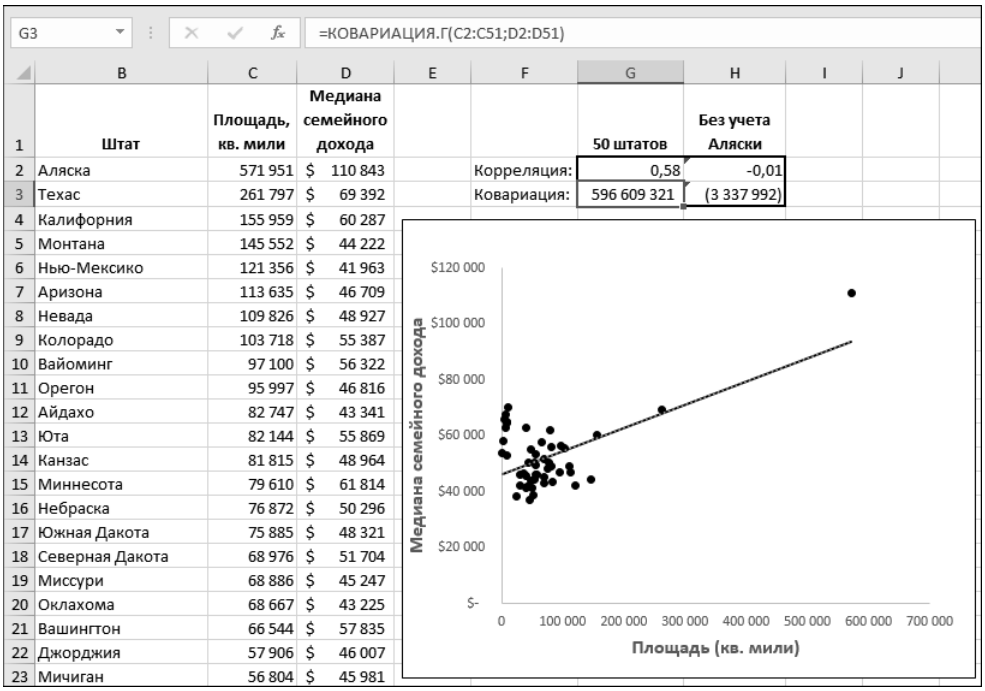

***Рис. 2.13.*** *Коэффициент корреляции Пирсона не позволяет получить точное количественное описание зависимости между данными переменными*

На рис. 2.13 приведена диаграмма зависимости между количеством тестовых манипуляций, точно выполняемых вручную в течение одной минуты (вертикальная

ось), и временем, отведенным для того, чтобы иметь возможность попрактиковаться в выполнении тестового задания (горизонтальная ось). Налицо пороговый эффект уменьшения полезной отдачи: начиная с некоторого значения дальнейшее увеличение длительности подготовки не приводит к улучшению результатов. Поэтому линия регрессии на рис. 2.13 искривлена, т.е. не является прямой.

Используя методики, о которых рассказывается в последующих главах, можно рассчитать показатели нелинейности зависимости между этими двумя переменными. В данном случае показатель нелинейной корреляции составляет 0,87, в то время как коэффициент корреляции Пирсона недооценивает степень зависимости, возвращая значение 0,79. Разница в степени корреляции 0,08, особенно в области возможных высоких значений коэффициента корреляции, довольно существенна.

Вы могли бы даже не заметить эту нелинейность, если бы не взглянули на график, приведенный на рис. 2.13. На рис. 2.14 проиллюстрирована схожая проблема, которая сразу же бросается в глаза, если отобразить данные на диаграмме.




| | B | C | D | E | F | G | H | I | J |
|---|---|---|---|---|---|---|---|---|---|
| 1 | Штат | Площадь, кв. мили | Медиана семейного дохода | | | 50 штатов | Без учета Аляски | | |
| 2 | Аляска | 571 951 | $ 110 843 | | Корреляция: | 0,58 | -0,01 | | |
| 3 | Техас | 261 797 | $ 69 392 | | Ковариация: | 596 609 321 | (3 337 992) | | |
| 4 | Калифорния | 155 959 | $ 60 287 | | | | | | |
| 5 | Монтана | 145 552 | $ 44 222 | | | | | | |
| 6 | Нью-Мексико | 121 356 | $ 41 963 | | | | | | |
| 7 | Аризона | 113 635 | $ 46 709 | | | | | | |
| 8 | Невада | 109 826 | $ 48 927 | | | | | | |
| 9 | Колорадо | 103 718 | $ 55 387 | | | | | | |
| 10 | Вайоминг | 97 100 | $ 56 322 | | | | | | |
| 11 | Орегон | 95 997 | $ 46 816 | | | | | | |
| 12 | Айдахо | 82 747 | $ 43 341 | | | | | | |
| 13 | Юта | 82 144 | $ 55 869 | | | | | | |
| 14 | Канзас | 81 815 | $ 48 964 | | | | | | |
| 15 | Миннесота | 79 610 | $ 61 814 | | | | | | |
| 16 | Небраска | 76 872 | $ 50 296 | | | | | | |
| 17 | Южная Дакота | 75 885 | $ 48 321 | | | | | | |
| 18 | Северная Дакота | 68 976 | $ 51 704 | | | | | | |
| 19 | Миссури | 68 886 | $ 45 247 | | | | | | |
| 20 | Оклахома | 68 667 | $ 43 225 | | | | | | |
| 21 | Вашингтон | 66 544 | $ 57 835 | | | | | | |
| 22 | Джорджия | 57 906 | $ 46 007 | | | | | | |
| 23 | Мичиган | 56 804 | $ 45 981 | | | | | | |

***Рис. 2.14.*** *Без построения диаграммы можно легко пропустить выброс*

Рис. 2.14 — это немного измененная версия рис. 2.2. Я добавил приблизительно `$50000` к медиане семейного дохода населения Аляски — изменение существенное, но именно такого рода вещи могут происходить из-за ошибок при вводе информации или неверно сформированных запросов к базам данных.

Значение коэффициента Пирсона `-0,01`, рассчитанное на основе данных о 49 штатах, указывает на отсутствие связи между семейным доходом и площадью штата, в котором проживает семья. Но если бы вы включили в расчет корреляции выброс, связанный

с одним из штатов, что вполне могло бы произойти, если бы он был глубоко скрыт где-то посередине списка, возвращенного запросом к базе данных, то получили бы коэффициент корреляции 0,58, соответствующий сильной зависимости между переменными.

Причиной столь разительного скачка величины коэффициента корреляции является увеличение ковариации от немногим более 3 миллионов до примерно 600 миллионов. В свою очередь, это обусловлено удаленностью значений медианы семейного дохода и площади для штата Аляска от средних значений для остальных штатов (оба этих значения, представленные на рис. 2.14, являются фиктивными и используются исключительно для того, чтобы с большей очевидностью проиллюстрировать эффекты, связанные с наличием выбросов в данных). Возведение этих отклонений в квадрат, что и происходит при включении их в вычисления, приводит к усилению корреляции.

Просматривая диаграммы, полезно искать не только признаки нелинейности зависимости между переменными, но и выбросы, благодаря чему вы будете точно знать, что именно происходит с вашими данными. Построение диаграмм на основе значений переменных, как на рис. 2.14, — не единственный способ диагностики проблем с данными, которые вы хотите анализировать с помощью корреляции. Однако эта методика позволяет обнаруживать не только нелинейность зависимости, но и явно ошибочные выбросы, а также другие аномалии. Excel настолько упрощает построение диаграмм, что отказываться от такой возможности было бы просто неразумно.

Для тех, кто никогда не имел дела с диаграммами в Excel, ниже приведено краткое описание того, как построить диаграмму, приведенную на рис. 2.14.

1. Выделите диапазон C2:C51, который содержит данные о площади каждого из штатов, выраженные в квадратных милях.
2. Удерживая нажатой клавишу <Ctrl>, выделите диапазон D2:D51, который содержит данные о среднем семейном доходе в каждом из штатов (ошибочные в случае Аляски).

**ПРИМЕЧАНИЕ**

Если переменные занимают смежные столбцы, как в случае столбцов C и D на рис. 2.14, можно выделить все столбцы сразу, а не по отдельности. В данном примере для этого достаточно щелкнуть на ячейке C2 и перетащить указатель мыши до ячейки D51.

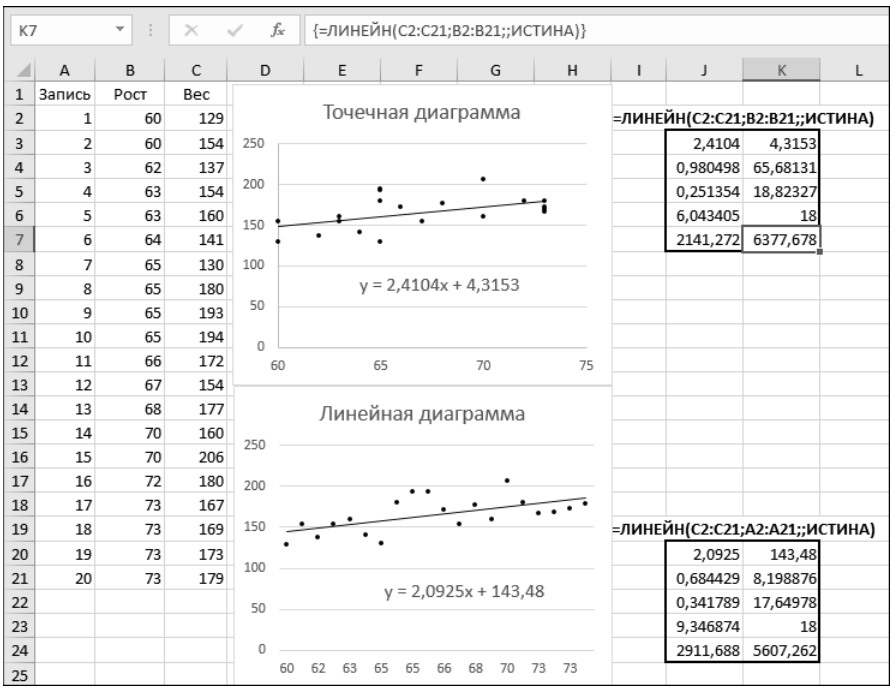
3. Щелкните на вкладке Вставка ленты и найдите группу элементов управления Диаграммы. Поочередно наводите указатель мыши на значки диаграмм различных типов, пока не найдете значок с текстом всплывающей подсказки “Вставить точечную (X, Y) или пузырьковую диаграмму”. Щелкните на этом значке.

Полученная точечная диаграмма в основном должна выглядеть так, как показано на рис. 2.14. Щелкните правой кнопкой мыши на любом маркере данных в области диаграммы и выберите в открывшемся контекстном меню пункт Добавить линию тренда. Примите предложенный по умолчанию вариант Линейная и закройте панель Формат линии тренда.

Немного подстроив параметры, в том числе удалив горизонтальные и вертикальные линии сетки, вы получите точную копию диаграммы, приведенной на рисунке.

Я не случайно рекомендовал вам выбрать вариант точечной диаграммы, поскольку (наряду с пузырьковым типом) это единственный тип диаграмм, в котором и горизонтальная, и вертикальная оси трактуются как оси значений. Excel различает оси значений и оси категорий. Ось значений представляет числовые величины и сохраняет количественные различия между ними. Так, расстояние от начала координат до точки на оси значений с координатой 8 в два раза превышает аналогичное расстояние до точки с координатой 4. В то же время расстояния между смежными точками 0, 8 и 9 на оси категорий будут одинаковыми: Excel трактует их как категории, и нет никаких причин для того, например, чтобы такие значения, как `Пепси`, `Кока-кола` и `Спрайт`, располагались на оси категорий не эквидистантно.

На рис. 2.15 приведен пример того, что произойдет, если построить не точечную, а, скажем, линейную диаграмму.




| | A | B | C |
|---|---|---|---|
| 1 | Запись | Рост | Вес |
| 2 | 1 | 60 | 129 |
| 3 | 2 | 60 | 154 |
| 4 | 3 | 62 | 137 |
| 5 | 4 | 63 | 154 |
| 6 | 5 | 63 | 160 |
| 7 | 6 | 64 | 141 |
| 8 | 7 | 65 | 130 |
| 9 | 8 | 65 | 180 |
| 10 | 9 | 65 | 193 |
| 11 | 10 | 65 | 194 |
| 12 | 11 | 66 | 172 |
| 13 | 12 | 67 | 154 |
| 14 | 13 | 68 | 177 |
| 15 | 14 | 70 | 160 |
| 16 | 15 | 70 | 206 |
| 17 | 16 | 72 | 180 |
| 18 | 17 | 73 | 167 |
| 19 | 18 | 73 | 169 |
| 20 | 19 | 73 | 173 |
| 21 | 20 | 73 | 179 |

***Рис. 2.15.*** *Линии регрессии кажутся одинаковыми, однако сравните между собой их уравнения*

Обе диаграммы, представленные на рис. 2.15 для демонстрации эффекта использования линейной диаграммы вместо точечной, отображают связь между двумя переменными, в данном случае переменными `Рост` и `Вес`. При построении точечной диаграммы обе переменные, изменяющиеся соответственно вдоль горизонтальной и вертикальной оси, трактуются как числовые. Для наблюдения в строке 2 переменная `Рост` имеет значение `60`, а переменная `Вес` — `129`.

В отличие от этого, при построении линейной диаграммы переменная, изменяющаяся вдоль вертикальной оси, трактуется как числовая, а переменная, изменяющаяся

вдоль горизонтальной оси, — как *категорийная* (*номинальная*). Коль скоро речь идет о линейной диаграмме, первые три значения переменной Рост вполне можно было бы обозначить как Алабама, Аляска и Аризона, а не 60, 60 и 62. (Я отформатировал ряд данных на линейной диаграмме таким образом, чтобы подавить отображение линий, которыми по умолчанию соединяются ее маркеры. Это было сделано для того, чтобы облегчить сопоставление позиций маркеров данных на диаграммах.)

Одним из следствий вышесказанного является отсутствие в Excel возможностей количественного различения наблюдений, откладываемых вдоль горизонтальной оси в подобных случаях. Заметьте, что на рис. 2.15 наблюдения, использованные для построения линейной диаграммы, располагаются вдоль горизонтальной оси эквидистантно. Надписи делений вдоль этой оси — это и есть всего лишь надписи, а не числовые значения. В то же время в случае точечной диаграммы расстояния между наблюдениями зависят от относительных значений переменной Рост.

Если запросить построение линии тренда, активизировав перед этим линейную диаграмму, то для расчета уравнения регрессии Excel потребуются числовые значения переменной Вес. Выделив линейную диаграмму и поместив на ее горизонтальную ось значения переменной Рост, вы сообщили Excel, что эти значения в действительности являются лишь названиями категорий. В связи с этим Excel использует для идентификации записей порядковые числа: программа интерпретирует значение переменной Рост в первой записи как 1, во второй записи — как 2 и т.д.

Отследим этот эффект, обратившись к результатам работы функции ЛИНЕЙН(), представленным в диапазоне J20:K24 на рис. 2.15. Эти результаты позволяют оценить степень зависимости между переменной Вес (столбец C) и номером записи (столбец A). Обратите внимание на то, что значения коэффициента регрессии и константы, хранящиеся в ячейках J20 и K20, совпадают с соответствующими значениями в уравнении регрессии.

Ознакомившись с аналогичными результатами для точечной диаграммы, отображенными в диапазоне J3:K7, вы увидите, что значения коэффициента регрессии и константы, возвращаемые функцией ЛИНЕЙН(), также совпадают с соответствующими значениями в уравнении регрессии. Однако в данном случае функция ЛИНЕЙН() оценивает степень зависимости между переменными Вес и Рост, а не между переменной Вес и номером записи.

Как следует из представленных данных, различия в трактовке осей значений и осей категорий в Excel влияют на то, как работает анализ диаграмм. Поскольку в выражении для коэффициента корреляции предполагается, что для измерения обеих переменных используются интервальные шкалы или шкалы отношений, для их графического представления следует всегда выбирать только диаграммы точечного типа, а не другие типы диаграмм, которые лишь выглядят как точечные. (Пузырьковые диаграммы также находят применение, однако, если речь идет о корреляционном анализе, они лишь запутывают картину.)

## Остерегайтесь ловушки при построении диаграмм

Коль скоро мы обсуждаем отображение данных с помощью точечных диаграмм, я хочу предостеречь вас от одной ловушки, над разгадкой которой мне пришлось

ломать голову целых два часа. Однако сначала я должен забежать немного наперед и затронуть определенные вопросы регрессионного анализа, о которых более подробно речь пойдет в главе 5, но которые целесообразно обсудить прямо сейчас в контексте построения диаграмм коррелирующих переменных.

Анализ корреляций играет важную роль в качестве подготовки к последующему регрессионному анализу, который может быть привлечен для прогнозирования значений одной переменной по известным значениям другой. Например, данные о корреляции между ростом и весом людей могут быть использованы для прогнозирования веса человека, если известен его рост.

Хотя я и не рекомендую использовать его в качестве вашего единственного средства для предсказания значений одной переменной по известному поведению другой, один из возможных способов получения прогнозного уравнения — построение точечной диаграммы (рис. 2.16).

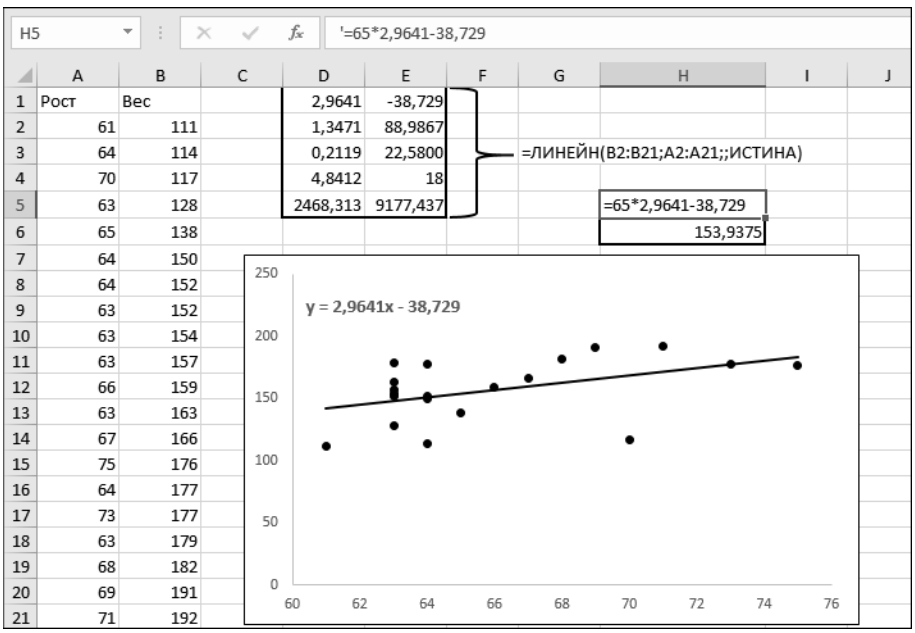



***Рис. 2.16.*** *По вашему запросу Excel рассчитает уравнение и отобразит его на диаграмме*

Диаграмма, приведенная на рис. 2.16, включает *уравнение регрессии*, которое позволяет прогнозировать значения переменной *Y* по известным значениям *X*. Excel получает это уравнение, используя информацию (и, в частности, коэффициент корреляции), касающуюся взаимозависимости переменных `Рост` и `Вес`, значения которых содержатся в диапазонах A2:A21 и B2:B21. Детали всех вычислений будут подробно рассмотрены в главах 3 и 4, а пока что вам достаточно знать, что данное уравнение предписывает выполнение следующих действий:

- выберите некоторое значение, например 65, для переменной Рост, которая в данном примере выступает в качестве предикторной (независимой, объясняющей);
- умножьте это значение на 2,9641;
- вычтите из полученного значения 38,729;
- полученный результат (153,9) представляет собой прогнозируемое значение веса человека, соответствующее значению роста 65.

А вот что необходимо сделать для того, чтобы получить данное уравнение с помощью диаграммы.

1. Создайте точечную диаграмму.
2. Щелкните правой кнопкой мыши на любом из маркеров данных на диаграмме.
3. Выберите в открывшемся контекстном меню пункт Добавить линию тренда.
4. Когда откроется панель Формат линии тренда, оставьте установленный по умолчанию переключатель Линейная включенным. Воспользуйтесь, если потребуется, полосой прокрутки панели, чтобы был виден флажок показывать уравнение на карте, и установите его.
5. Закройте панель Формат линии тренда.

В результате выполнения этих действий на диаграмме отобразится уравнение регрессии. В случае необходимости перетащите уравнение в другое место, если маркер данных или какой-либо другой элемент управления закрывает его.

Это же уравнение можно получить другим способом, который почти всегда оказывается более удобным: с помощью функции ЛИНЕЙН(). Здесь я лишь кратко коснусь этого вопроса, поскольку рассмотрению функции ЛИНЕЙН() и ее возможностей посвящена глава 4. Результаты применения функции ЛИНЕЙН() к данным, содержащимся в столбцах A и B, отображаются в диапазоне D1:E5 (см. рис. 2.16). Обратите внимание на значения ячеек D1 и E1: они совпадают с соответствующими значениями в уравнении регрессии. Функция ЛИНЕЙН() выполняет, в частности, те же действия, которые предписываются уравнением на диаграмме: умножает значение переменной Рост на число 2,9641 и вычитает из полученного результата числа 38,729 (на самом деле она добавляет число -38,729, но это эквивалентная операция). Поэтому, если вы хотите предсказать вес человека, рост которого составляет 65 дюймов, используйте следующее уравнение:

$$y = 2{,}9641x - 38{,}729$$

У вас может возникнуть вопрос: откуда Excel известно, какую из переменных следует считать независимой, т.е. $x$, а какую — прогнозируемой, т.е. $y$, при построении точечной диаграммы? Разработчики Excel давным-давно договорились о том, что переменная, которая находится на рабочем листе слева от другой переменной, должна изменяться вдоль горизонтальной оси. Переменная, которая появляется справа от другой переменной, должна рассматриваться как переменная, изменяющаяся вдоль вертикальной оси.

Вы можете ввести значения двух переменных в двух несмежных столбцах и воспользоваться их групповым выделением (выделить значения первой переменной, нажать и удерживать клавишу <Ctrl>, а затем выделить значения второй переменной). Результат будет тем же — вдоль горизонтальной оси будут откладываться значения той переменной, которая находится слева.

Кроме того, в соответствии с принятым в Excel соглашением значения переменной, которая прогнозируется с помощью уравнения регрессии, отображаемого на диаграмме, откладываются вдоль вертикальной оси. Переменная, которой на диаграмме соответствует горизонтальная ось, рассматривается как независимая в уравнении регрессии.

Поэтому, если переменной `Рост` соответствует горизонтальная ось, то смысл уравнения регрессии

```
Y = 2,9641(X) – 38,729
```

заключается в том, что для получения прогнозного значения переменной `Y` (вес) следует умножить значение переменной `X` (рост) на число `2,9641` и вычесть из полученного результата число `38,729`.

Резюмируя, можно сказать, что в уравнении регрессии, отображаемом на диаграмме, прогнозируемой считается та переменная, значения которой расположены на рабочем листе **справа** от значений другой переменной.

Возвращаясь к рис. 2.16, считаю важным обратить ваше внимание на способ задания аргументов в выражении вызова функции `ЛИНЕЙН()`, введенном в ячейках D1:E5:

```
=ЛИНЕЙН(B2:B21;A2:A21;;ИСТИНА)
```

Синтаксис вызова функции `ЛИНЕЙН()` предполагает явное указание ролей переменных их позициями при передаче в качестве аргументов и не учитывает взаимное расположение значений переменных на рабочем листе. В приведенной выше формуле предполагается, что значения прогнозируемой переменной содержатся в ячейках B2:B21. Эти значения всегда указываются в качестве **первого** аргумента функции `ЛИНЕЙН()`. Значения независимой переменной находятся в ячейках A2:A21. Эти значения всегда указываются в качестве **второго** аргумента функции `ЛИНЕЙН()`.

Описанный порядок расстановки переменных — определение независимой переменной по ее позиции на рабочем листе при вычислении уравнения регрессии на диаграмме и по ее позиции в списке аргументов при вызове функции рабочего листа — не совсем удачен, поскольку его вряд ли можно считать последовательным. Однако у тех, кто его продумывал, были, вероятно, свои соображения для принятия такого решения, и мы должны с этим мириться.

И все же разработчики упустили из виду склонность людей (включая меня) совершать ошибки в силу инерционности мышления. Обратимся к рис. 2.17.

Наиболее очевидное различие между рис. 2.16 и 2.17 заключается в том, что на рис. 2.16 прогнозируемая переменная, `Вес`, занимает столбец B, а на рис. 2.17 — столбец A.

Взгляните на результаты, возвращенные функцией `ЛИНЕЙН()` в диапазоне D1:E1. Если вы сравните эти значения с формулой, которая отображается на диаграмме, то увидите, что коэффициент (ячейка D1) и константа (ячейка E1) уже не совпадают с соответствующими значениями в уравнении.

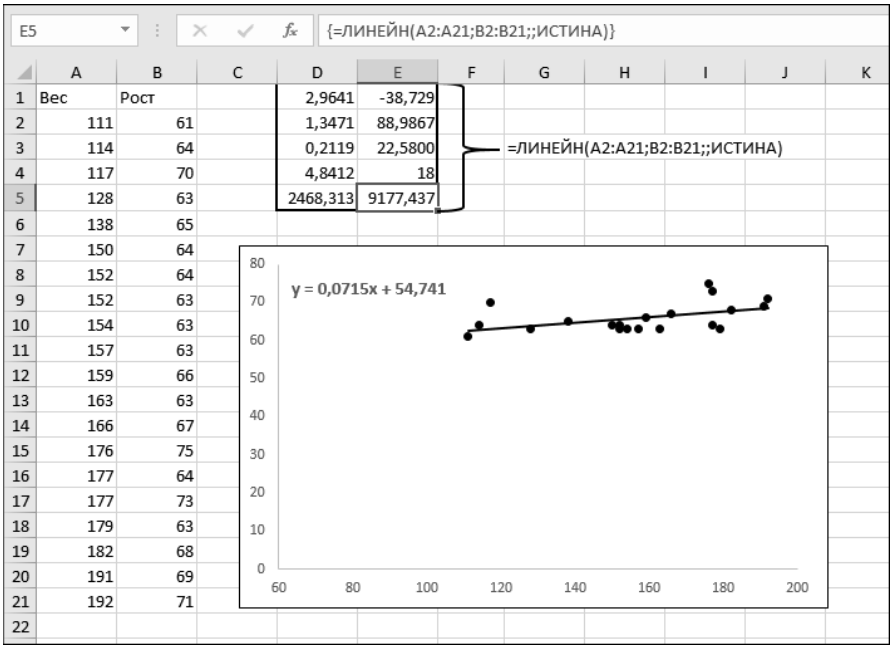



***Рис. 2.17.*** *Результаты работы функции* `ЛИНЕЙН()` *не согласуются с приведенным на диаграмме уравнением регрессии*

Это непосредственно объясняется тем фактом, что, как на рис. 2.16, так и на рис. 2.17, переменная `Вес` трактуется функцией `ЛИНЕЙН()` как прогнозируемая. Прогнозируемая переменная всегда передается функции `ЛИНЕЙН()` в качестве первого аргумента. На рис. 2.16 таким аргументом является диапазон B2:B21, на рис. 2.17 — диапазон A2:A21.

Однако, поскольку на рис. 2.17 переменная `Вес` заняла позицию справа от переменной `Рост`, на диаграмме ей соответствует горизонтальная ось, а переменной `Рост` — вертикальная. Затем, когда рассчитываются линия регрессии и регрессионная формула, прогнозируемой переменной считается не `Вес`, а `Рост`.

Результаты описаны ниже:

- маркеры данных на рис. 2.17 развернуты на 90 градусов относительно их позиций на рис. 2.16;
- изменилось расположение линии регрессии;
- изменилось уравнение регрессии, отображаемое на диаграмме.

В свое время аналогичная ситуация сбила меня с толку. Я никак не мог понять, почему коэффициент и константа, вычисленные с помощью функции `ЛИНЕЙН()`, не совпадают с коэффициентом и константой в уравнении регрессии, отображаемом на диаграмме. Значительная часть выполняемой мною работы так или иначе связана с регрессионным анализом, в связи с чем были все основания думать, что я должен немедленно определить источник проблемы и устранить ее. Наверное, от меня действительно следовало ожидать, что в подобной ситуации я окажусь на высоте.

Согласен, так и должно было произойти. Но хочу заметить, что необходимость в отображении уравнения регрессии на диаграмме возникает в моей практике крайне редко, и поэтому оно встречалось мне при подобных обстоятельствах не так уж часто. На протяжении многих лет мне просто не приходилось сталкиваться с непоследовательностью описанного подхода.

Но как могло случиться, что я поместил прогнозируемую переменную в столбец A, а независимую переменную — в ячейку B? Я сделал это машинально, интуитивно следуя очередности задания аргументов функции `ЛИНЕЙН()`, и рассказываю об этом столь подробно лишь для того, чтобы показать вам, как легко можно допустить оплошность, если заранее не знаешь, где тебя подстерегает ловушка.

Функция `ЛИНЕЙН()` способна предсказывать значения одиночных переменных, таких как `Вес`, на основе значений не только одной, но и *нескольких* независимых переменных. (Этот аспект обсуждается в главе 5.) Чтобы использовать функцию `ЛИНЕЙН()` с несколькими независимыми переменными, их следует размещать в смежных столбцах. Поэтому, если вы предполагаете, что в дальнейшем вам могут понадобиться дополнительные независимые переменные, резервируйте для них свободное место на рабочем листе, как показано в примере на рис. 2.18.

| | A | B | C | D | E |
|---|---|---|---|---|---|
| 1 | Рост | Вес | | | |
| 2 | 61 | 111 | | | |
| 3 | 64 | 114 | | | |
| 4 | 70 | 117 | | | |
| 5 | 63 | 128 | | | |
| 6 | 65 | 138 | | | |
| 7 | 64 | 150 | | | |
| 8 | 64 | 152 | | | |
| 9 | 63 | 152 | | | |
| 10 | 63 | 154 | | | |
| 11 | 63 | 157 | | | |
| 12 | 66 | 159 | | | |
| 13 | 63 | 163 | | | |
| 14 | 67 | 166 | | | |
| 15 | 75 | 176 | | | |
| 16 | 64 | 177 | | | |
| 17 | 73 | 177 | | | |
| 18 | 63 | 179 | | | |
| 19 | 68 | 182 | | | |
| 20 | 69 | 191 | | | |
| 21 | 71 | 192 | | | |

| | A | B | C | D |
|---|---|---|---|---|
| 1 | Вес | Рост | Возраст | Еженедельная длительность тренировок (минуты) |
| 2 | 111 | 61 | 13 | 55 |
| 3 | 114 | 64 | 20 | 86 |
| 4 | 117 | 70 | 36 | 101 |
| 5 | 128 | 63 | 47 | 120 |
| 6 | 138 | 65 | 18 | 59 |
| 7 | 150 | 64 | 67 | 66 |
| 8 | 152 | 64 | 54 | 38 |
| 9 | 152 | 63 | 15 | 66 |
| 10 | 154 | 63 | 76 | 70 |
| 11 | 157 | 63 | 20 | 33 |
| 12 | 159 | 66 | 16 | 59 |
| 13 | 163 | 63 | 96 | 38 |
| 14 | 166 | 67 | 57 | 77 |
| 15 | 176 | 75 | 32 | 73 |
| 16 | 177 | 64 | 69 | 51 |
| 17 | 177 | 73 | 74 | 48 |
| 18 | 179 | 63 | 16 | 118 |
| 19 | 182 | 68 | 89 | 109 |
| 20 | 191 | 69 | 19 | 32 |
| 21 | 192 | 71 | 81 | 24 |

***Рис. 2.18.*** *При добавлении на исходный рабочий лист (слева) дополнительных независимых переменных (справа) может потребоваться перестановка данных*

Предположим, ваши данные расположены так, как показано на рис. 2.18, *слева*. Проанализировав соотношение между переменными `Рост` и `Вес`, вы решаете перейти к множественному регрессионному анализу и вводите дополнительную переменную `Возраст`. Чтобы включить переменную `Возраст` в анализ, вам нужно добавить для нее новый столбец, вставив его, например, между столбцами переменных `Рост` и `Вес`, чтобы независимые переменные `Рост` и `Возраст` оказались в смежных столбцах.

Наиболее разумно поместить прогнозируемую переменную в крайний левый столбец, как показано на рис. 2.18, *справа*. Это позволит легко достраивать набор смежных столбцов для дополнительных независимых переменных, если в этом возникнет необходимость. Однако учтите, что при построении диаграмм Excel будет интерпретировать прогнозируемую переменную, оказавшуюся в крайнем слева столбце, как независимую.

Я чувствую себя немного неловко из-за того, что посвятил несколько страниц описанию проблемы, с которой вам, возможно, никогда не придется столкнуться. Имейте, однако, в виду, что вам вовсе не обязательно запрашивать отображение уравнения регрессии на диаграмме, и это может ввести вас в заблуждение относительно того, какая из переменных является прогнозируемой, а какая — независимой. Положение и наклон линии регрессии определяются уравнением регрессии. Поэтому, прежде чем запрашивать вывод линии тренда, убедитесь в том, что вам точно известно, что именно представляет собой диаграмма.

И если десять минут, которые вы потратили на чтение истории о том, как я споткнулся при работе с диаграммой Excel, избавят вас в будущем от пары часов головной боли, то этот рассказ не был напрасным.

## Корреляция и причинно-следственная связь

Вероятно, вам приходилось читать или слышать о различии между корреляцией и каузальными (причинно-следственными) отношениями: несмотря на то что корреляция часто означает наличие причинно-следственной связи, она не может служить доказательством того, что так оно и есть.

Предположим, в ходе своих исследований вы обнаружили, что для многих участков восточного и западного побережий США наблюдается положительная корреляция между средней температурой океанской воды в пределах километровой прибрежной зоны и количеством китов, зафиксированных в сезон их миграции.

Кое-кто мог бы выдвинуть предположение о том, что теплая вода привлекает организмы, которыми питаются киты, и поэтому киты плывут туда, где есть пища. И если коэффициент корреляции оказался достаточно большим, скажем 0,80, чтобы вы в соответствии со своими критериями сочли корреляционную связь сильной, то вы могли бы прийти к выводу, что такое объяснение является корректным.

Не дайте ввести себя в заблуждение эмпирическими свидетельствами и кажущейся логичностью такой аргументации. Корреляционный анализ ничего не доказывает. Статистика не используется для демонстрации того, истинна или ложна теория. Для исключения конкурирующих объяснений результатов наблюдений ставят *плановые эксперименты*. Статистика же привлекается для обобщения информации, собранной в ходе таких экспериментов, и количественной оценки вероятности того, что принимаемое решение может быть неверным при имеющейся доказательной базе.

Обычно в тех случаях, когда пытаются объяснить наблюдения причинно-следственной связью, выдвигают два конкурирующих типа гипотез. Гипотезы первого типа касаются направленности предполагаемой причинно-следственной связи; при этом причина может быть ошибочно истолкована как следствие. В гипотезах второго типа предполагается наличие одной или нескольких дополнительных переменных, которые оказывают влияние на обе исходные переменные.

## Направление причинно-следственной связи

Было бы преувеличением утверждать, что тепло, исходящее от тела кита, служит причиной повышения температуры океанской воды. Однако корреляционная зависимость между количеством владельцев огнестрельного оружия и количеством убийств с применением огнестрельного оружия уже можно рассматривать в качестве реалистичного примера, иллюстрирующего направленность причинно-следственных связей. Можно менять методологии эксперимента, но все они будут давать значение коэффициента корреляции между количеством огнестрельного оружия, находящегося во владении граждан на территории административной единицы (муниципалитета, штата, страны), и числом убийств, совершенных этим оружием, приблизительно равное 0,30.

Если в основе этого соотношения лежит причинно-следственная связь, то что является причиной, а что следствием? Вызывает ли увеличение количества имеющегося у населения оружия рост числа убийств? А может быть, увеличение количества единиц приобретаемого оружия является следствием реакции людей на повышенный уровень преступности в данном регионе?

Здесь мы не будем пытаться дать ответы на эти вопросы. На самом деле никто и не собирается получать ответы на них посредством исследования корреляционных зависимостей. Сама по себе корреляция не в состоянии продемонстрировать ни наличие причинно-следственной связи, ни, если она имеется, ее направление. Экспериментальный подход должен включать случайный выбор большого количества разных регионов и последующее случайное отнесение их к одной из двух групп. В одной группе от жителей потребовали бы приобретать больше оружия. В другой группе жителям запретили бы приобретение дополнительного оружия.

По прошествии некоторого времени следовало бы подсчитать число убийств с использованием огнестрельного оружия, учитывая поправку на ковариацию с количеством оружия, находившегося во владении граждан до начала эксперимента. Если результаты, полученные для двух указанных групп, не будут существенно различаться, то гипотеза о том, что увеличение количества имеющегося оружия ведет к росту числа убийств, должна быть отброшена.

Совершенно очевидно, что в силу различных обстоятельств, включая причины этического характера, правовые аспекты и вообще осуществимость, такой эксперимент никогда не будет проведен. Поэтому многие исследователи вынуждены прибегать к корреляционному анализу, подменяя им подлинные плановые эксперименты. Мне часто попадаются отчеты о результатах исследований, в которых содержатся предположения о существовании связи между интенсивностью использования мобильных телефонов и частотой раковых заболеваний, вакцинацией и аутизмом, расстоянием от жилья до линий электропередачи и (опять-таки) раковыми заболеваниями и т.п. В одних случаях исследователи отталкиваются от непосредственных характеристик изучаемого поведения (таких, например, как частота использования мобильного телефона), тогда как в других используют корреляционные коэффициенты (и их кажущуюся статистическую значимость).

Мы можем и должны использовать корреляционный анализ, поскольку он облегчает обнаружение возможных закономерностей, которые подлежат последующей проверке путем проведения более строгих целевых экспериментов. Если между событиями наблюдается корреляция, то это дает основания предполагать, что между ними существует причинно-следственная связь и, возможно, корреляция обусловлена объективными причинами. Однако сам по себе факт наличия корреляции еще не является доказательством того, что такая связь действительно существует.

Если в силу каких-либо экономических, юридических или этических соображений проведение подлинных экспериментов невозможно, то все, что нам остается делать, это продолжать наблюдения за интересующими нас событиями в надежде, что по прошествии достаточно длительного времени проблема прояснится. В 50–60-х годах прошлого столетия считалось, что одной из причин рака легких и горла является курение. В качестве аргумента в свое оправдание представители табачной индустрии указывали на отсутствие надежных экспериментальных данных, способных убедительно доказать, что между курением и раковыми заболеваниями существует прямая причинно-следственная связь. Они также подчеркивали тот факт, что различные корреляционные исследования так и не смогли доказать вину табака: корреляция не тождественна причинно-следственной связи.

Однако с тех пор прошло много лет, и с учетом накопленных за это время результатов дополнительных корреляционных исследований вряд ли кто-то сегодня будет безоговорочно отрицать тот факт, что курение способствует возникновению раковых заболеваний.

Очевидно одно: если сама по себе корреляция еще не означает обязательное наличие причинно-следственной связи, то достаточно сильная корреляция может служить убедительным аргументом в пользу этого.

Между 1930 и 1936 годами в Ольденбурге (Германия) наблюдался резкий рост численности населения. За тот же период было замечено значительно возросшее число прилетов аистов. В соответствии с данным тогда шутливым объяснением этого факта непосредственной причиной роста численности населения города стало увеличение количества детей, приносимых аистами. Разумеется, если принять, что опубликованные данные были достоверными, то направление причинно-следственной связи было обратным: при большем количестве населения шансы наблюдать прилеты аистов увеличиваются.

## Третья переменная

Иногда с обеими переменными, вовлеченными в исследуемую корреляцию, оказывается связанной причинными связями третья переменная. Например, до середины 1990-х годов бытовало мнение об отсутствии устойчивого соотношения между количеством полицейских на душу населения и уровнем преступности. Однако полученные с тех пор результаты более строгих исследований и статистического анализа показали, что такое соотношение действительно существует.

В то же время, о чем говорят результаты недавних корреляционных исследований, существование описанной корреляции не обязательно указывает на ее причинно-следственную природу. Как на численность полицейских, так и на уровень преступности оказывает влияние третья переменная: социально-экономические условия,

существующие в конкретном сообществе. Как правило, в более богатых общинах налоги на доходы больше, что позволяет тратить на содержание полиции более значительные суммы. Для этих сообществ характерен более низкий уровень преступности и, в частности, меньшее количество преступлений, совершенных с применением насилия.

Результаты этих исследований говорят о том, что простое увеличение численности полицейских еще не приводит к снижению количества преступлений. Соотношения между различными переменными носят сложный характер и трудно поддаются разложению на отдельные составляющие, но полученные результаты дают основания полагать, что благосостояние сообщества оказывает влияние как на количество полицейских, так и на уровень преступности.

## Ограничение диапазона

Дополнительный момент, о котором следует помнить, — это потеря полного диапазона значений одной из переменных.

По крайней мере начиная с 1990-х годов стандартные тесты наподобие SAT[1] подвергаются критике по целому ряду причин разного характера. В частности, экзамены, сдаваемые при поступлении в колледж, критикуют за то, что их результаты не позволяют надежно судить о возможной фактической успеваемости студентов в будущем.

Если бы вы взяли выборку, состоящую из 50 студентов колледжа, и сопоставили их средние баллы в процессе обучения с их оценками, полученными при сдаче SAT, то, вероятно, получили бы результат, близкий к тому, который представлен на рис. 2.19. Несмотря на принципиальную возможность получения абитуриентами более низких оценок SAT, из того факта, что этим 50 студентов удалось закончить колледж, следует, что в данной фиктивной выборке нижняя граница оценок SAT равна приблизительно 1100.

Точки данных на диаграмме расположены в основном случайным образом, но характер их расположения указывает на существование умеренной корреляции между оценками SAT и средними баллами в процессе обучения, что находит свое отражение в значении коэффициента корреляции 0,40.

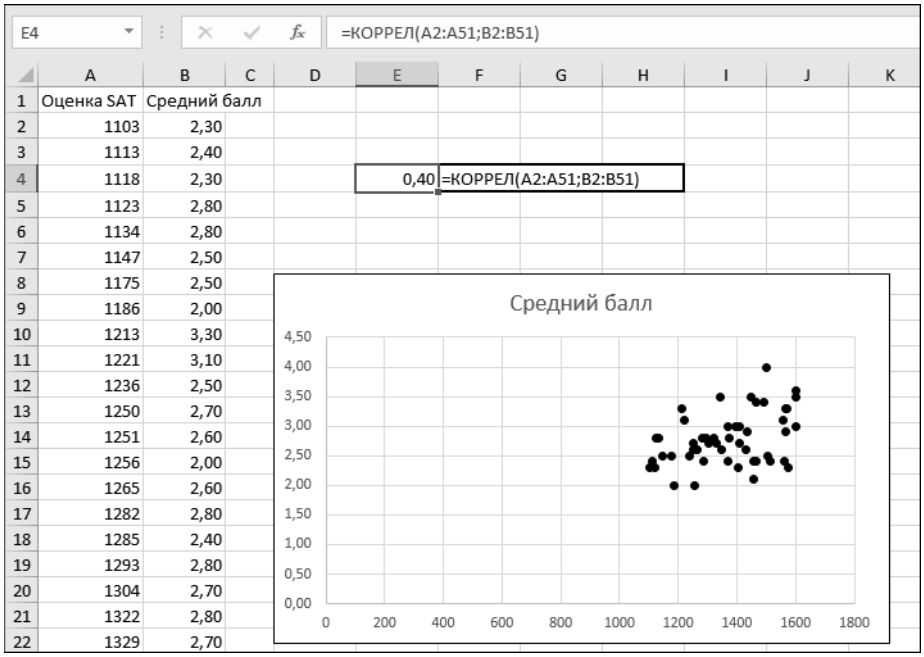
Если бы это были все доступные вам данные, то вы могли бы сделать вывод, что оценки SAT плохо предсказывают успеваемость студентов колледжа.

А теперь предположим, что у вас имеется доступ к данным о других 50 студентах, которые вынуждены были преждевременно покинуть колледж из-за низкой успеваемости (рис. 2.20). Поскольку они не закончили колледж, их данные не были включены в анализ, результаты которого приведены на рис. 2.19.

Характер расположения маркеров данных на рис. 2.19 и 2.20 весьма сходен, однако на рис. 2.20 как оценки SAT, так и средние баллы студентов находятся в диапазонах более низких значений. И опять-таки, по рисунку нельзя сказать, что оценки SAT позволяют делать заключения об успеваемости студентов.

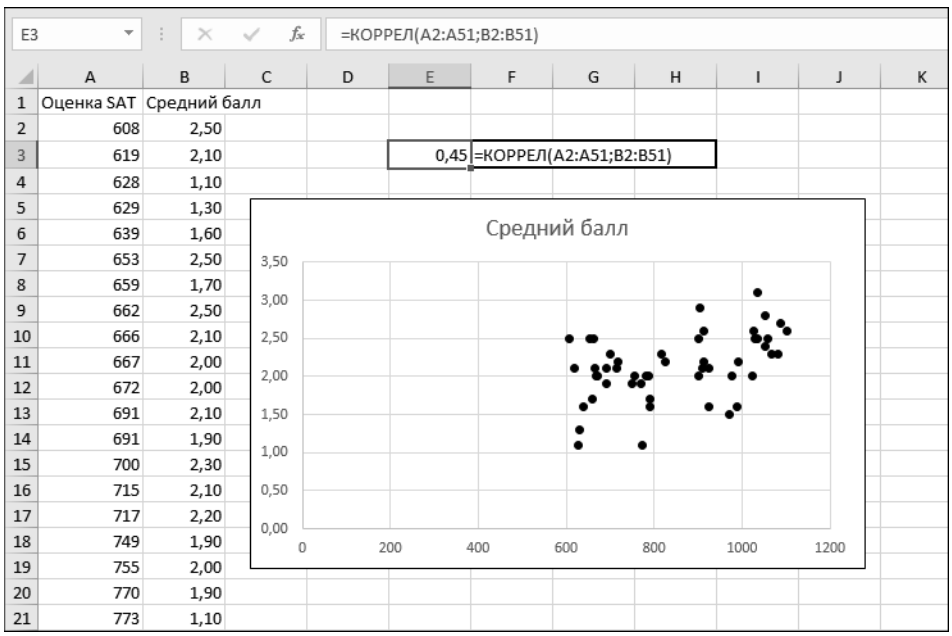
Но что произойдет, если мы поместим все 100 пар оценок на одну диаграмму? Именно это и сделано на рис. 2.21.

[1] SAT (Scholastic Assessment Test — академический оценочный тест) — стандартизированный тест для приема в высшие учебные заведения в США. — *Примеч. ред.*




| | A | B |
|---|---|---|
| 1 | Оценка SAT | Средний балл |
| 2 | 1103 | 2,30 |
| 3 | 1113 | 2,40 |
| 4 | 1118 | 2,30 |
| 5 | 1123 | 2,80 |
| 6 | 1134 | 2,80 |
| 7 | 1147 | 2,50 |
| 8 | 1175 | 2,50 |
| 9 | 1186 | 2,00 |
| 10 | 1213 | 3,30 |
| 11 | 1221 | 3,10 |
| 12 | 1236 | 2,50 |
| 13 | 1250 | 2,70 |
| 14 | 1251 | 2,60 |
| 15 | 1256 | 2,00 |
| 16 | 1265 | 2,60 |
| 17 | 1282 | 2,80 |
| 18 | 1285 | 2,40 |
| 19 | 1293 | 2,80 |
| 20 | 1304 | 2,70 |
| 21 | 1322 | 2,80 |
| 22 | 1329 | 2,70 |

***Рис. 2.19.*** *Коэффициент корреляции 0,40 не является свидетельством сильной взаимосвязи*




| | A | B |
|---|---|---|
| 1 | Оценка SAT | Средний балл |
| 2 | 608 | 2,50 |
| 3 | 619 | 2,10 |
| 4 | 628 | 1,10 |
| 5 | 629 | 1,30 |
| 6 | 639 | 1,60 |
| 7 | 653 | 2,50 |
| 8 | 659 | 1,70 |
| 9 | 662 | 2,50 |
| 10 | 666 | 2,10 |
| 11 | 667 | 2,00 |
| 12 | 672 | 2,00 |
| 13 | 691 | 2,10 |
| 14 | 691 | 1,90 |
| 15 | 700 | 2,30 |
| 16 | 715 | 2,10 |
| 17 | 717 | 2,20 |
| 18 | 749 | 1,90 |
| 19 | 755 | 2,00 |
| 20 | 770 | 1,90 |
| 21 | 773 | 1,10 |

***Рис. 2.20.*** *Коэффициент корреляции 0,45 указывает лишь на чуть более сильную корреляцию между переменными, чем коэффициент 0,40 (см. рис. 2.19)*

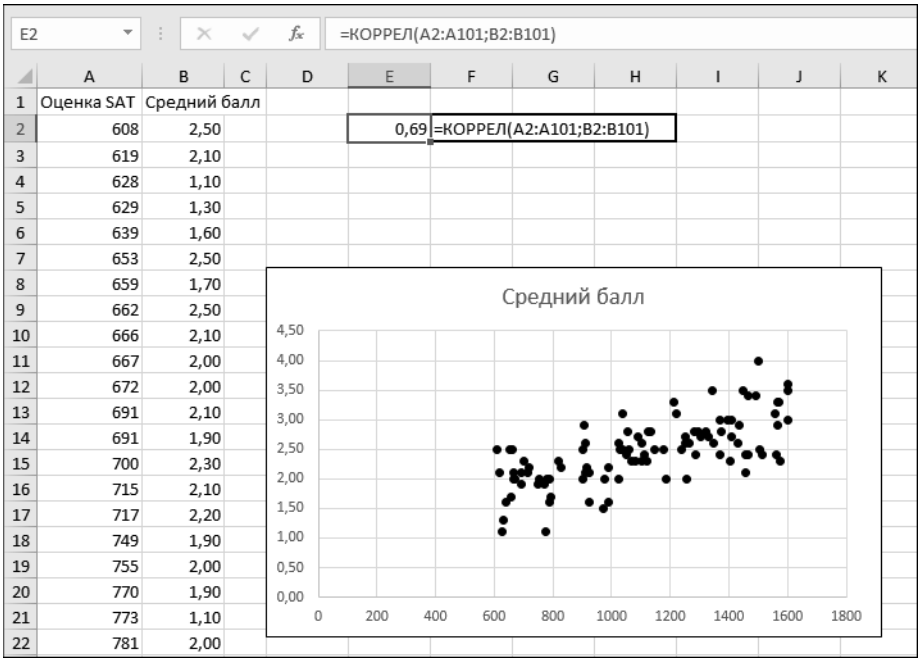



***Рис. 2.21.*** *После снятия ограничений на диапазоны возможных значений переменных коэффициент корреляции почти достиг отметки 0,70*

На рис. 2.19 и 2.20 ограничения, наложенные на диапазоны изменения обеих переменных — оценки SAT и среднего балла, — искусственно занижали значение коэффициента корреляции, характеризующего кажущуюся степень зависимости переменных. Несмотря на то что утверждение о довольно слабой степени корреляции между оценками SAT и средними баллами студентов, оставшихся в колледже, было бы корректным, оно значительно отличается от утверждения о слабой корреляции этих переменных для всех студентов в целом, из чего следовало бы, что SAT — не совсем подходящий инструмент для отбора студентов. Хорошо это или плохо, но результаты SAT используются для распределения дефицитного ресурса — количества набираемых студентов. Было бы неверно считать корреляцию слабой, если она оказалась таковой лишь в результате того, что диапазон исходных данных для расчета коэффициента корреляции был искусственно сужен.

В главе 1 обсуждались понятия и методы, имеющие отношение к дисперсии. В данной главе рассматривались вопросы, связанные с ковариацией, т.е. взаимосогласованным изменением значений двух переменных или отсутствием такового. В главе 3, посвященной простой регрессии, весь материал, который мы к этому времени успели рассмотреть, послужит основой для того, чтобы показать, как можно использовать полученные знания для более глубокого изучения данных, с которыми вы регулярно работаете. Также будет показано, что совместное рассмотрение вариации и корреляции случайных переменных образует фундамент для более сложных методов, таких как множественная регрессия, факторный анализ дисперсии и ковариационный анализ.