Looking at Relationships

2

We think of a dataset as a collection of d-tuples (a d-tuple is an ordered list of d elements). For example, the Chase and Dunner dataset had entries for Gender; Grade; Age; Race; Urban/Rural; School; Goals; Grades; Sports; Looks; and Money (so it consisted of 11-tuples). The previous chapter explored methods to visualize and summarize a set of values obtained by extracting a single element from each tuple. For example, I could visualize the heights or the weights of a population (as in Fig. 1.7). But I could say nothing about the relationship between the height and weight. In this chapter, we will look at methods to visualize and summarize the relationships between pairs of elements of a dataset.

2.1 Plotting 2D Data

We take a dataset, choose two different entries, and extract the corresponding elements from each tuple. The result is a dataset consisting of 2-tuples, and we think of this as a two dimensional dataset. The first step is to plot this dataset in a way that reveals relationships. The topic of how best to plot data fills many books, and we can only scratch the surface here. Categorical data can be particularly tricky, because there are a variety of choices we can make, and the usefulness of each tends to depend on the dataset and to some extent on one's cleverness in graphic design (Sect. 2.1.1).

For some continuous data, we can plot the one entry as a function of the other (so, for example, our tuples might consist of the date and the number of robberies; or the year and the price of lynx pelts; and so on, Sect. 2.1.2).

Mostly, we use a simple device, called a scatter plot. Using and thinking about scatter plots will reveal a great deal about the relationships between our data items (Sect. 2.1.3).

2.1.1 Categorical Data, Counts, and Charts

Categorical data is a bit special. Assume we have a dataset with several categorical descriptions of each data item. One way to plot this data is to think of it as belonging to a richer set of categories. Assume the dataset has categorical descriptions, which are not ordinal. Then we can construct a new set of categories by looking at each of the cases for each of the descriptions. For example, in the Chase and Dunner data of Table 1.2, our new categories would be: "boy-sports"; "girl-sports"; "boy-popular"; "girl-popular"; "boy-grades"; and "girl-grades". A large set of categories like this can result in a poor bar chart, though, because there may be too many bars to group the bars successfully. Figure 2.1 shows such a bar chart. Notice that it is hard to group categories by eye to compare; for example, you can see that slightly more girls think grades are important than boys do, but to do so you need to compare two bars that are separated by two other bars. An alternative is a **pie chart**, where a circle is divided into sections whose angle is proportional to the size of the data item. You can think of the circle as a pie, and each section as a slice of pie. Figure 2.1 shows a pie chart, where each section is proportional to the number of students in its category. In this case, I've used my judgement to lay the categories out in a way that makes comparisons easy. I'm not aware of any tight algorithm for doing this, though.

Pie charts have problems, because it is hard to judge small differences in area accurately by eye. For example, from the pie chart in Fig. 2.1, it's hard to tell that the "boy-sports" category is slightly bigger than the "boy-popular" category (try it;

D. Forsyth, *Probability and Statistics for Computer Science*,
https://doi.org/10.1007/978-3-319-64410-3_2

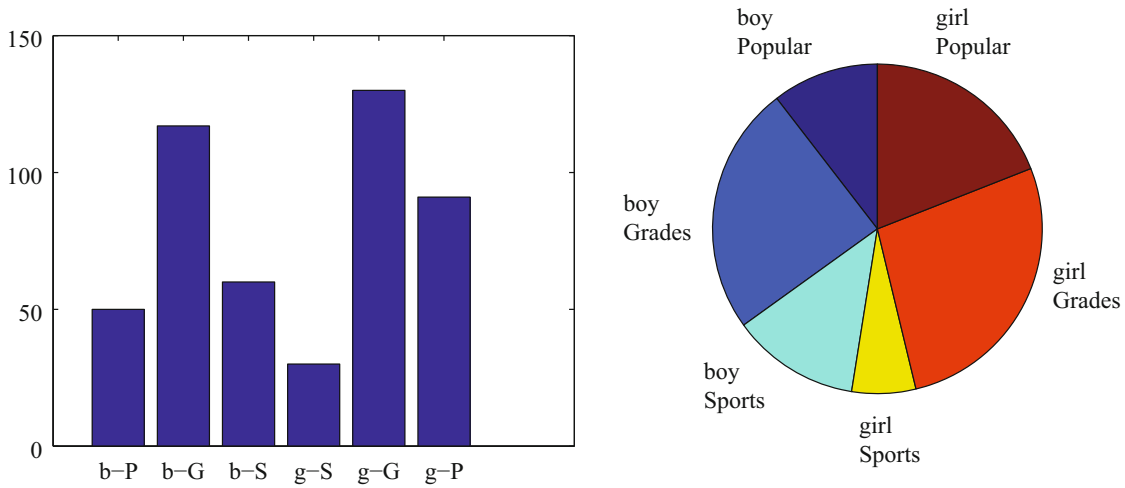

Fig. 2.1 I sorted the children in the Chase and Dunner study into six categories (two genders by three goals), and counted the number of children that fell into each cell. I then produced the bar chart on the *left*, which shows the number of children of each gender, selecting each goal. On the *right*, a pie chart of this information. I have organized the pie chart so it is easy to compare boys and girls by eye—start at the top; going down on the left side are boy goals, and on the right side are girl goals. Comparing the size of the corresponding wedges allows you to tell what goals boys (resp. girls) identify with more or less often

check using the bar chart). For either kind of chart, it is quite important to think about *what* you plot. For example, the plot of Fig. 2.1 shows the total number of respondents, and if you refer to Fig. 1.1, you will notice that there are slightly more girls in the study. Is the *percentage* of boys who think grades are important smaller (or larger) than the *percentage* of girls who think so? you can't tell from these plots, and you'd have to plot the percentages instead.

An alternative is to use a **stacked bar chart**. You can (say) regard the data as of two types, "Boys" and "Girls". Within those types, there are subtypes ("Popularity", "Grades" and "Sport"). The height of the bar is given by the number of elements in the type, and the bar is divided into sections corresponding to the number of elements of that subtype. Alternatively, if you want the plot to show relative frequencies, the bars could all be the same height, but the shading corresponds to the fraction of elements of that subtype. This is all much harder to say than to see or to do (Fig. 2.2).

An alternative to a pie chart that is very useful for two dimensional data is a **heat map**. This is a method of displaying a matrix as an image. Each entry of the matrix is mapped to a color, and the matrix is represented as an image. For the Chase and Dunner study, I constructed a matrix where each row corresponds to a choice of "sports", "grades", or "popular", and each column corresponds to a choice of "boy" or "girl". Each entry contains the count of data items of that type. Zero values are represented as white; the largest values as red; and as the value increases, we use an increasingly saturated pink. This plot is shown in Fig. 2.3

If the categorical data is ordinal, the ordering offers some hints for making a good plot. For example, imagine we are building a user interface. We build an initial version, and collect some users, asking each to rate the interface on scales for "ease of use" $(-2, -1, 0, 1, 2$, running from bad to good) and "enjoyability" (again, $-2, -1, 0, 1, 2$, running from bad to good). It is natural to build a 5×5 table, where each cell represents a pair of "ease of use" and "enjoyability" values. We then count the number of users in each cell, and build graphical representations of this table. One natural representation is a **3D bar chart**, where each bar sits on its cell in the 2D table, and the height of the bars is given by the number of elements in the cell. Table 2.1 shows a table and Fig. 2.4 shows a 3D bar chart for some simulated data. The main difficulty with a 3D bar chart is that some bars are hidden behind others. This is a regular nuisance. You can improve things by using an interactive tool to rotate the chart to get a nice view, but this doesn't always work. Heatmaps don't suffer from this problem (Fig. 2.4), another reason they are a good choice.

Remember this: *There are a variety of tools for plotting categorical data. It's difficult to give strict rules for which to use when, but usually one tries to avoid pie charts (angles are hard to judge by eye) and 3D bar charts (where occlusion can hide important effects).*

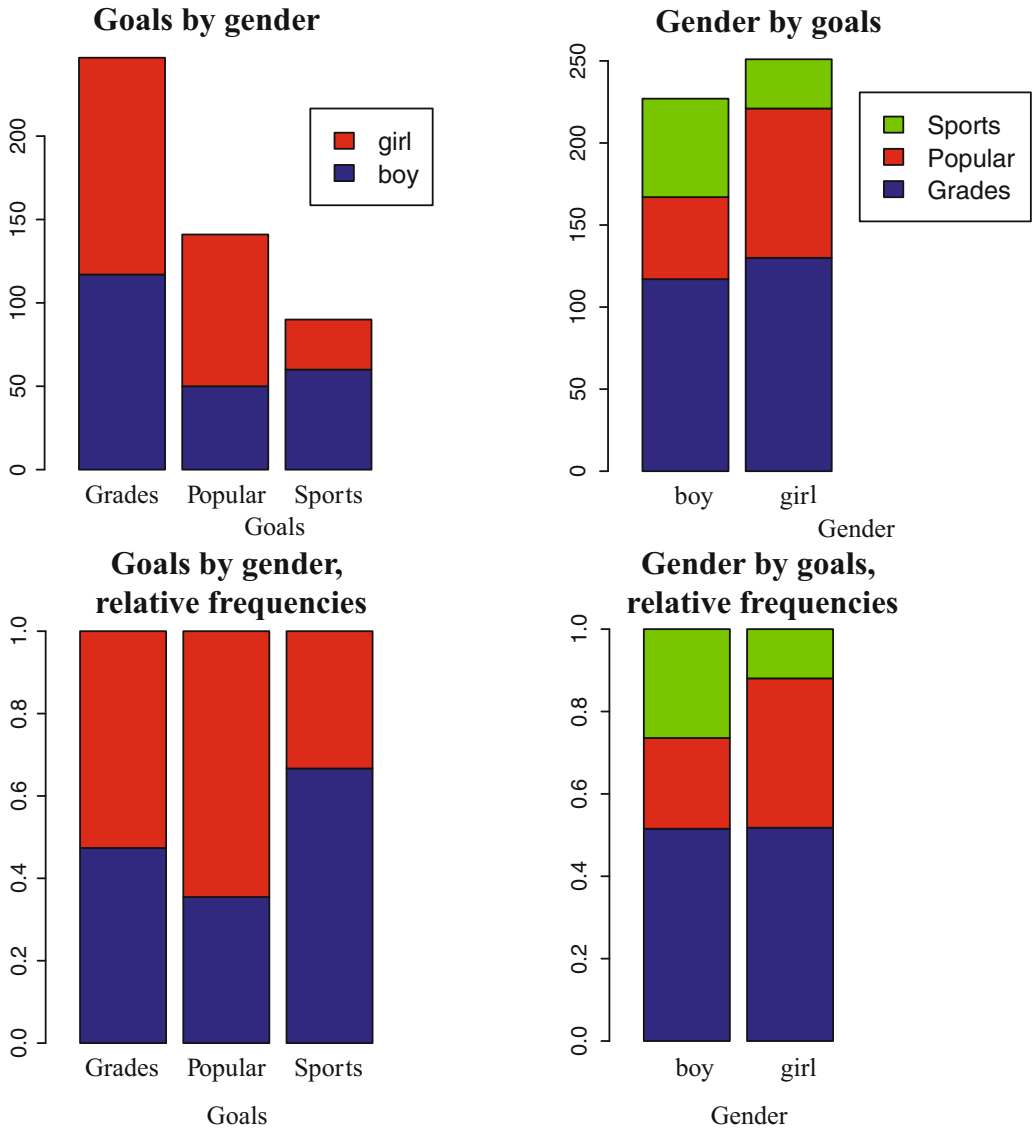

Fig. 2.2 These bar charts use stacked bars. In the *top row*, the overall height of the bar is given by the number of elements of that type but each different subtype is identified by shading, so you can tell by eye, for example, how many of the "Grades" in the study were "Boys". This layout makes it hard to tell what *fraction* of, say, "Boys" aspire to "Popularity". In the *bottom row*, all bars have the same height, but the shading of the bar identifies the fraction of that type that has a corresponding subtype. This means you can tell by eye what fraction of, for example, "Girls" aspire to "Sports"

2.1.2 Series

Sometimes one component of a dataset gives a natural ordering to the data. For example, we might have a dataset giving the maximum rainfall for each day of the year. We could record this either by using a two-dimensional representation, where one dimension is the number of the day and the other is the temperature, or with a convention where the i'th data item is the rainfall on the i'th day. For example, at http://lib.stat.cmu.edu/DASL/Datafiles/timeseriesdat.html, you can find four datasets indexed in this way. It is natural to plot data like this as a function of time. From this dataset, I extracted data giving the number of burglaries each month in a Chicago suburb, Hyde Park. I have plotted part this data in Fig. 2.5 (I left out the data to do with treatment effects). It is natural to plot a graph of the burglaries as a function of time (in this case, the number of the month). The plot shows each data point explicitly. I also told the plotting software to draw lines joining data points, because burglaries do not all happen on a specific day. The lines suggest, reasonably enough, the rate at which burglaries are happening between data points.

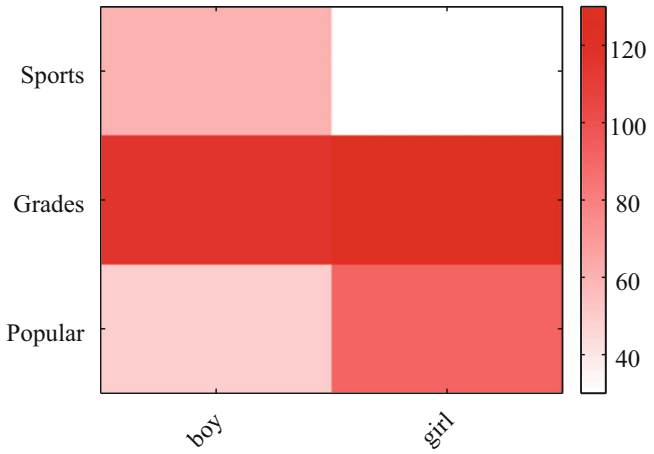

Fig. 2.3 A heat map of the Chase and Dunner data. The color of each cell corresponds to the count of the number of elements of that type. The colorbar at the side gives the correspondence between color and count. You can see at a glance that the number of boys and girls who prefer grades is about the same; that about the same number of boys prefer sports and popularity, with sports showing a mild lead; and that more girls prefer popularity to sports

Table 2.1 I simulated data representing user evaluations of a user interface

	−2	−1	0	1	2
−2	24	5	0	0	1
−1	6	12	3	0	0
0	2	4	13	6	0
1	0	0	3	13	2
2	0	0	0	1	5

Each cell in the table on the *left* contains the count of users rating "ease of use" (horizontal, on a scale of −2—very bad—to 2—very good) vs. "enjoyability" (vertical, same scale). Users who found the interface hard to use did not like using it either. While this data is categorical, it's also ordinal, so that the order of the cells is determined. It wouldn't make sense, for example, to reorder the columns of the table or the rows of the table

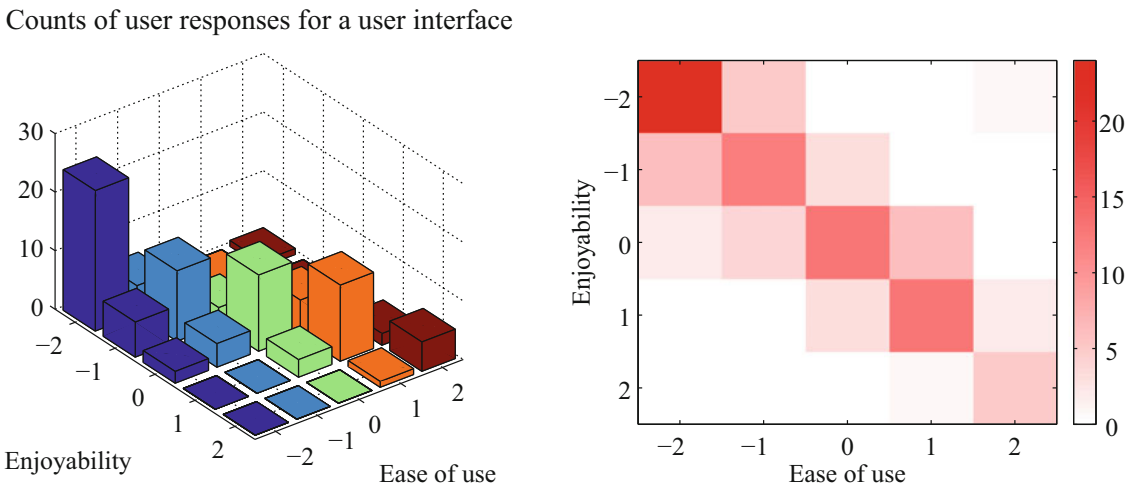

Fig. 2.4 On the *left*, a 3D bar chart of the data. The height of each bar is given by the number of users in each cell. This figure immediately reveals that users who found the interface hard to use did not like using it either. However, some of the bars at the back are hidden, so some structure might be hard to infer. On the *right*, a heat map of this data. Again, this figure immediately reveals that users who found the interface hard to use did not like using it either. It's more apparent that everyone disliked the interface, though, and it's clear that there is no important hidden structure

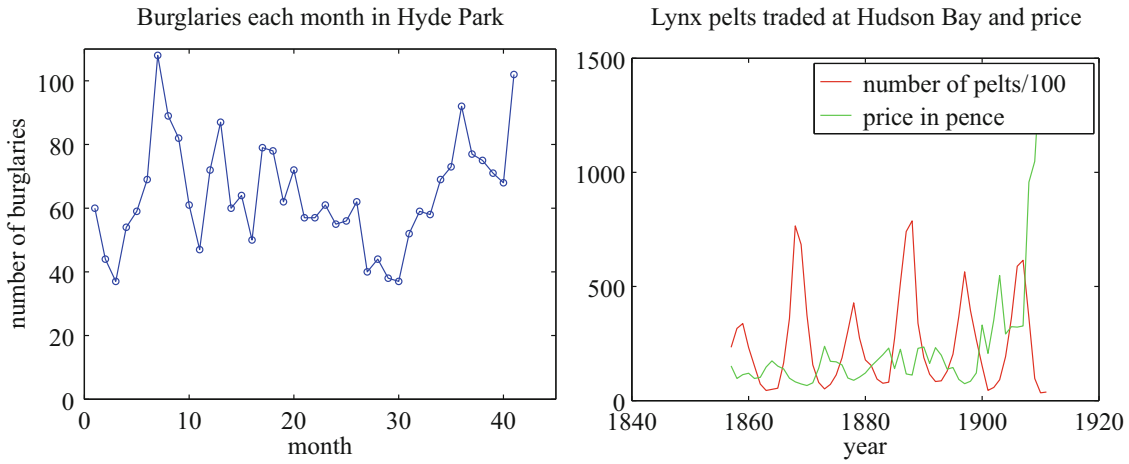

Fig. 2.5 *Left*, the number of burglaries in Hyde Park, by month. *Right*, a plot of the number of lynx pelts traded at Hudson Bay and of the price paid per pelt, as a function of the year. Notice the scale, and the legend box (the number of pelts is scaled by 100)

As another example, at http://lib.stat.cmu.edu/datasets/Andrews/ you can find a dataset that records the number of lynx pelts traded to the Hudson's Bay company and the price paid for each pelt. This version of the dataset appeared first in Table 3.2 of *Data: a Collection of Problems from many Fields for the Student and Research Worker* by D.F. Andrews and A.M. Herzberg, published by Springer in 1985. I have plotted it in Fig. 2.5. The dataset is famous, because it shows a periodic behavior in the number of pelts (which is a good proxy for the number of lynx), which is interpreted as a result of predator-prey interactions. Lynx eat rabbits. When there are many rabbits, lynx kittens thrive, and soon there will be many lynx; but then they eat most of the rabbits, and starve, at which point the rabbit population rockets. You should also notice that after about 1900, prices seem to have gone up rather quickly. I don't know why this is. There is also some suggestion, as there should be, that prices are low when there are many pelts, and high when there are few.

2.1.3 Scatter Plots for Spatial Data

It isn't always natural to plot data as a function. For example, in a dataset containing the temperature and blood pressure of a set of patients, there is no reason to believe that temperature is a function of blood pressure, or the other way round. Two people could have the same temperature, and different blood pressures, or vice-versa. As another example, we could be interested in what causes people to die of cholera. We have data indicating *where* each person died in a particular outbreak. It isn't helpful to try and plot such data as a function.

The **scatter plot** is a powerful way to deal with this situation. In the first instance, assume that our data points actually describe points on the a real map. Then, to make a scatter plot, we make a mark on the map at a place indicated by each data point. What the mark looks like, and how we place it, depends on the particular dataset, what we are looking for, how much we are willing to work with complex tools, and our sense of graphic design.

Figure 2.6 is an extremely famous scatter plot, due to John Snow. Snow—one of the founders of epidemiology—used a scatter plot to reason about a cholera outbreak centered on the Broad Street pump in London in 1854. At that time, the mechanism that causes cholera was not known. Snow plotted cholera deaths as little bars (more bars, more deaths) on the location of the house where the death occurred. More bars means more deaths, fewer bars means fewer deaths. There are more bars per block close to the pump, and few far away. This plot offers quite strong evidence of an association between the pump and death from cholera. Snow used this scatter plot as evidence that cholera was associated with water, and that the Broad Street pump was the source of the tainted water.

Remember this: *Scatter plots are a most effective tool for geographic data and 2D data in general. A scatter plot should be your first step with a new 2D dataset.*

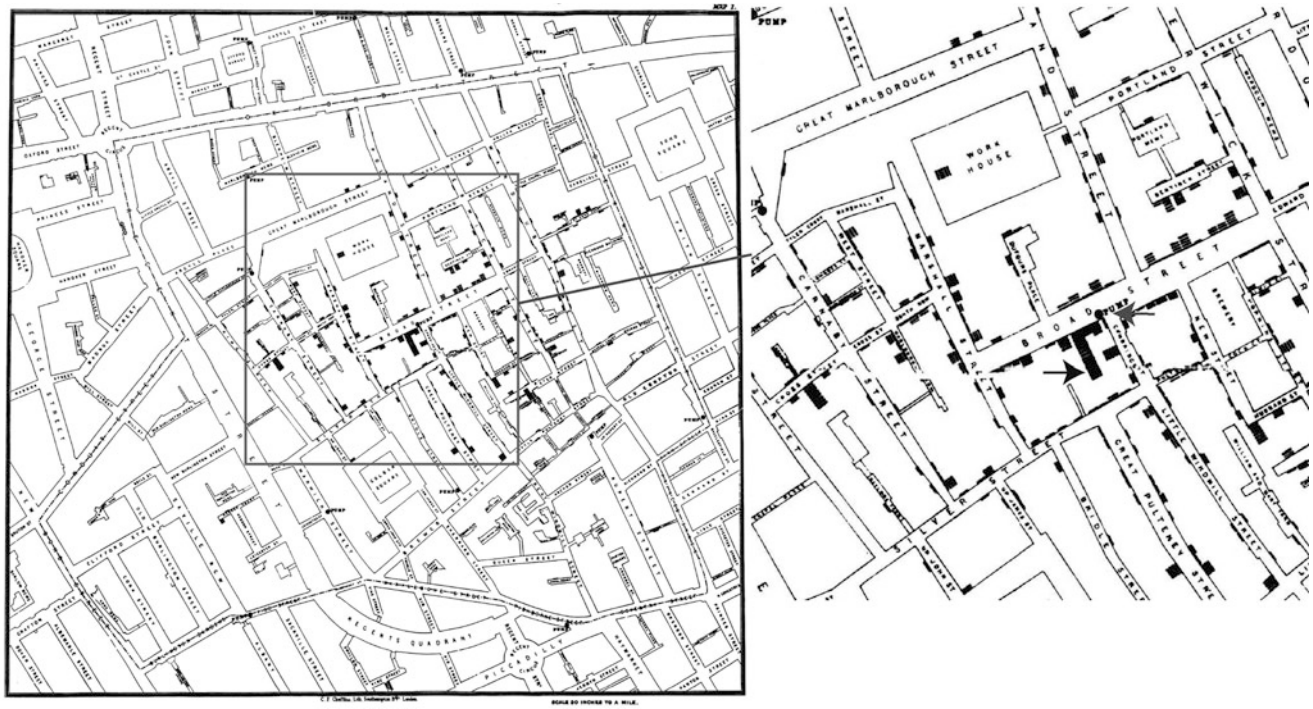

Fig. 2.6 Snow's scatter plot of cholera deaths on the *left*. Each cholera death is plotted as a small bar on the house in which the bar occurred (for example, the black arrow points to one stack of these bars, indicating many deaths, in the detail on the *right*). Notice the fairly clear pattern of many deaths close to the Broad street pump (grey arrow in the detail), and fewer deaths further away (where it was harder to get water from the pump)

2.1.4 Exposing Relationships with Scatter Plots

A scatter plot is a useful, simple tool for ferreting out associations in data. Now we need some notation. Assume we have a dataset $\{\mathbf{x}\}$ of N data items, $\mathbf{x}_1, \ldots, \mathbf{x}_N$. Each data item is a d dimensional vector (so its components are numbers). We wish to investigate the relationship between two components of the dataset. For example, we might be interested in the 7'th and the 13'th component of the dataset. We will produce a two-dimensional plot, one dimension for each component. It does not really matter which component is plotted on the x-coordinate and which on the y-coordinate (though it will be some pages before this is clear). But it is very difficult to write sensibly without talking about the x and y coordinates.

We will make a two-dimensional dataset out of the components that interest us. We must choose which component goes first in the resulting 2-vector. We will plot this component on the x-coordinate (and we refer to it as the x-coordinate), and to the other component as the y-coordinate. This is just to make it easier to describe what is going on; there's no important idea here. It really will not matter which is x and which is y. The two components make a dataset $\{\mathbf{x}_i\} = \{(x_i, y_i)\}$. To produce a scatter plot of this data, we plot a small shape at the location of each data item.

Such scatter plots are very revealing. For example, Fig. 2.7 shows a scatter plot of body temperature against heart rate for humans. In this dataset, the gender of the subject was recorded (as "1" or "2"—I don't know which is which), and so I have plotted a "1" at each data point with gender "1", and so on. Looking at the data suggests there isn't much difference between the blob of "1" labels and the blob of "2" labels, which suggests that females and males are about the same in this respect.

The scale used for a scatter plot matters. For example, plotting lengths in meters gives a very different scatter from plotting lengths in millimeters. Figure 2.8 shows two scatter plots of weight against height. Each plot is from the same dataset, but one is scaled so as to show two outliers. Keeping these outliers means that the rest of the data looks quite concentrated, just because the axes are in large units. In the other plot, the axis scale has changed (so you can't see the outliers), but the data looks more scattered. This may or may not be a misrepresentation. Figure 2.9 compares the data with outliers removed, with the same plot on a somewhat different set of axes. One plot looks as though increasing height corresponds to increasing weight; the other looks as though it doesn't. This is purely due to deceptive scaling—each plot shows the same dataset.

Dubious data can also contribute to scaling problems. Recall that, in Fig. 2.5, price data before and after 1900 appeared to behave differently. Figure 2.10 shows a scatter plot of the lynx data, where I have plotted number of pelts against price. I

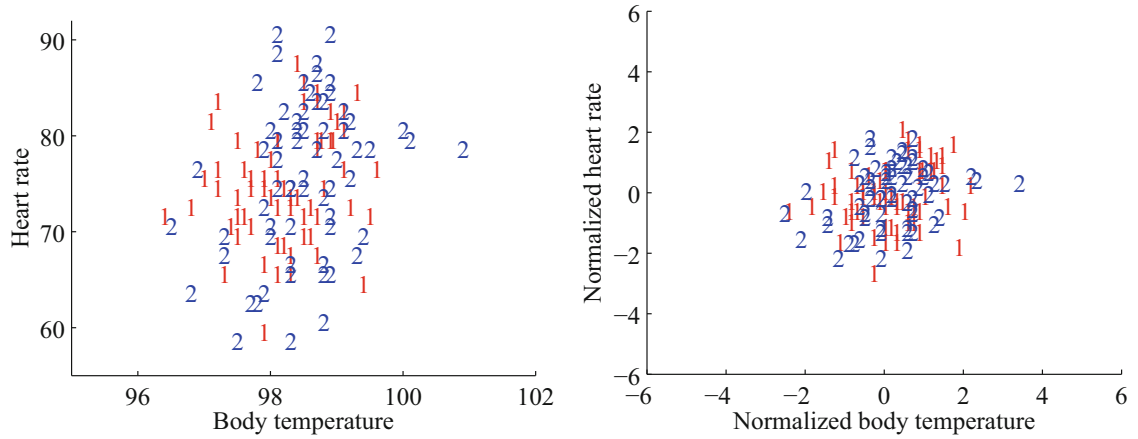

Fig. 2.7 A scatter plot of body temperature against heart rate, from the dataset at http://www2.stetson.edu/~jrasp/data.htm; normtemp.xls. I have separated the two genders by plotting a different symbol for each (though I don't know which gender is indicated by which letter); if you view this in color, the differences in color makes for a greater separation of the scatter. This picture suggests, but doesn't conclusively establish, that there isn't much dependence between temperature and heart rate, and any dependence between temperature and heart rate isn't affected by gender

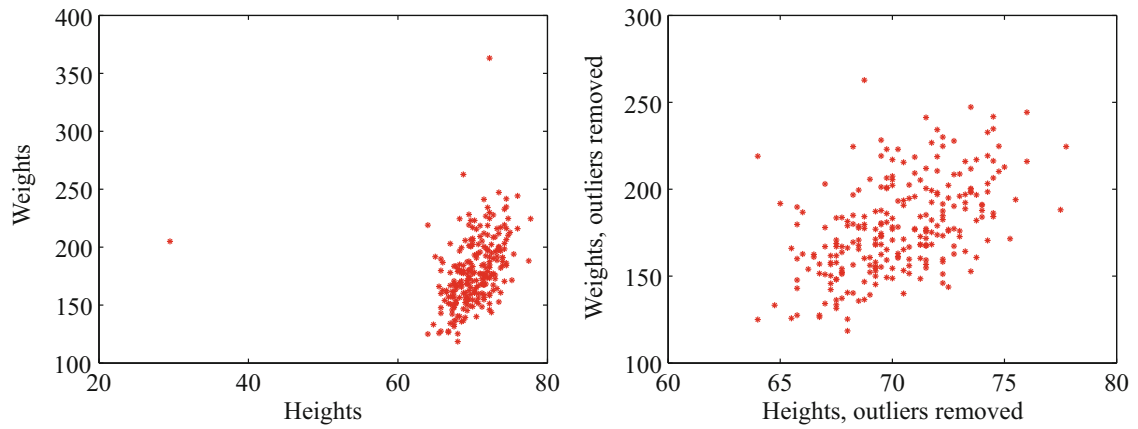

Fig. 2.8 Scatter plots of weight against height, from the dataset at http://www2.stetson.edu/~jrasp/data.htm. *Left:* Notice how two outliers dominate the picture, and to show the outliers, the rest of the data has had to be bunched up. *Right* shows the data with the outliers removed. The structure is now somewhat clearer

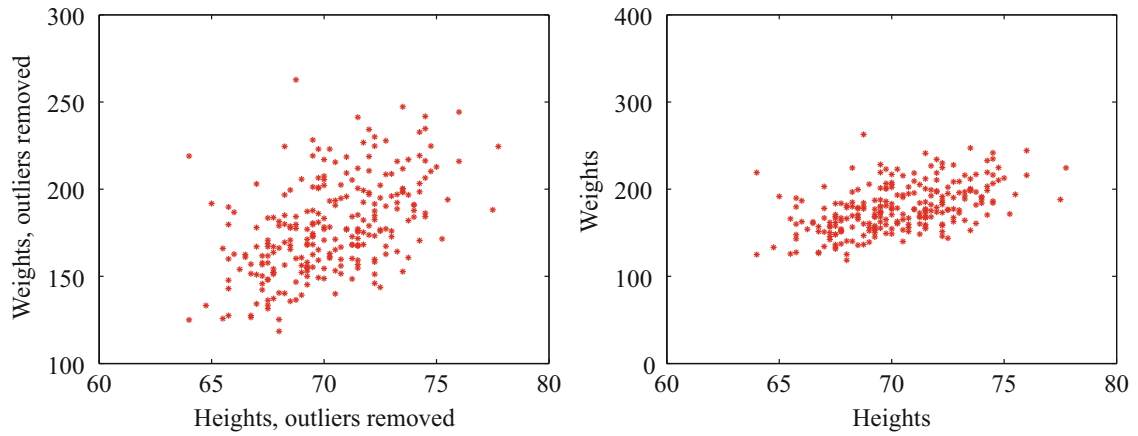

Fig. 2.9 Scatter plots of weight against height, from the dataset at http://www2.stetson.edu/~jrasp/data.htm. *Left:* data with two outliers removed, as in Fig. 2.8. *Right:* this data, rescaled slightly. Notice how the data looks less spread out. But there is no difference between the datasets. Instead, your eye is easily confused by a change of scale

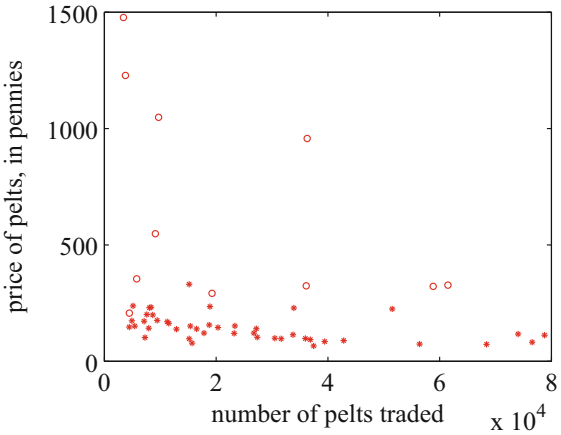

Fig. 2.10 A scatter plot of the price of lynx pelts against the number of pelts. I have plotted data for 1901 to the end of the series as circles, and the rest of the data as *'s. It is quite hard to draw any conclusion from this data, because the scale is confusing. Furthermore, the data from 1900 on behaves quite differently from the other data

plotted the post-1900 data as circles, and the rest as asterisks. Notice how the circles seem to form a quite different figure, which supports the suggestion that something interesting happened around 1900. We can reasonably choose to analyze data after 1900 separately from before 1900. A choice like this should be made with care. If you exclude every data point that might disagree with your hypothesis, you may miss the fact that you are wrong. Leaving out data is an essential component of many kinds of fraud. You should always reveal whether you have excluded data, and why, to allow the reader to judge the evidence.

When you look at Fig. 2.10, you should notice the scatter plot does not seem to support the idea that prices go up when supply goes down. This is puzzling because it's generally a pretty reliable idea. In fact, the plot is just hard to interpret because it is poorly scaled. Scale is an important nuisance, and it's easy to get misled by scale effects.

The way to avoid the problem is to plot in standard coordinates. We can normalize without worrying about the dimension of the data—we normalize each dimension independently by subtracting the mean of that dimension and dividing by the standard deviation of that dimension. This means we can normalize the x and y coordinates of the two-dimensional data separately. We continue to use the convention of writing the normalized x coordinate as $\hat{x}$ and the normalized y coordinate as $\hat{y}$. So, for example, we can write $\hat{x}_j = (x_j - \mathsf{mean}\,(\{x\}))/\mathsf{std}\,(\{x\}))$ for the $\hat{x}$ value of the j'th data item in normalized coordinates. Normalizing shows us the dataset on a standard scale. Once we have done this, it is quite straightforward to read off simple relationships between variables from a scatter plot.

Remember this: *The plot scale can mask effects in scatter plots, and it's usually a good idea to plot in standard coordinates.*

2.2 Correlation

Plotting data in standard coordinates can be very revealing. For example, it is pretty clear from Fig. 2.11 that someone who is taller than the mean will tend to be heavier than the mean too. This relationship isn't in the form of a function. There are some people who are quite a lot taller than the mean, and quite a lot lighter, too. But taller people are mostly heavier, too. There isn't always a relationship, as Fig. 2.12 suggests. There really doesn't seem to be any reason to suspect that heart rate and temperature are related. Sometimes the relationship goes the other way, i.e. when one variable increases, another decreases. Figure 2.13 strongly suggests that when more pelts were traded, the price tended to be lower.

The simplest, and most important, relationship to look for in a scatter plot is this: when $\hat{x}$ increases, does $\hat{y}$ tend to increase, decrease, or stay the same? This is straightforward to spot in a normalized scatter plot, because each case produces a very clear shape on the scatter plot. Any relationship is called **correlation** (we will see later how to measure this), and the three cases are: positive correlation, which means that larger $\hat{x}$ values tend to appear with larger $\hat{y}$ values; zero correlation, which means no relationship; and negative correlation, which means that larger $\hat{x}$ values tend to appear with smaller $\hat{y}$ values. You should notice that this relationship isn't a function—the data forms blobs, rather than lying on curves—and it isn't affected by swapping $\hat{x}$ and $\hat{y}$. If larger $\hat{x}$ tends to occur with larger $\hat{y}$, then larger $\hat{y}$ tends to occur with larger $\hat{x}$, and so on. Figure 2.14 compares a plot of height against weight to one of weight against height. Usually, one just does this by rotating the page,

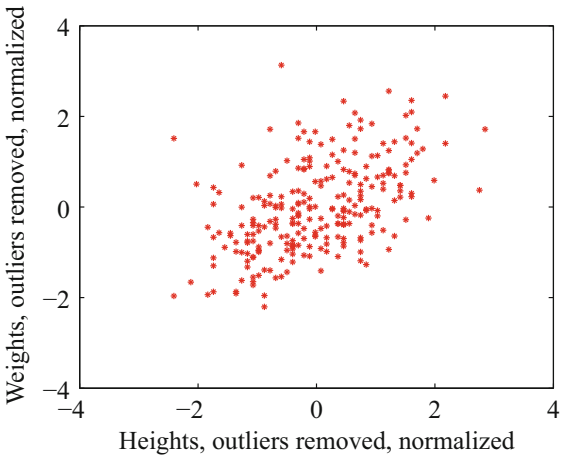

Fig. 2.11 A normalized scatter plot of weight against height, from the dataset at http://www2.stetson.edu/~jrasp/data.htm. Now you can see that someone who is a standard deviation taller than the mean will tend to be somewhat heavier than the mean too

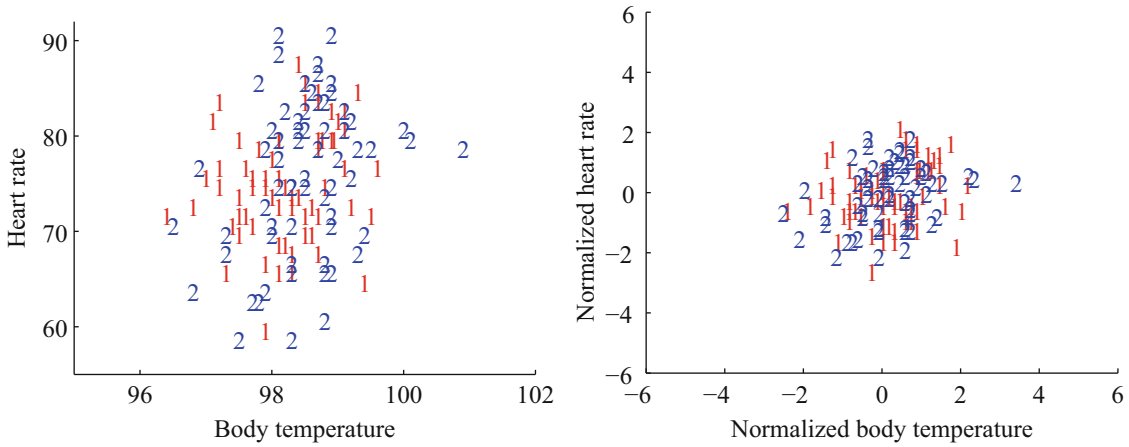

Fig. 2.12 *Left:* A scatter plot of body temperature against heart rate, from the dataset at http://www2.stetson.edu/~jrasp/data.htm; normtemp.xls. I have separated the two genders by plotting a different symbol for each (though I don't know which gender is indicated by which letter); if you view this in color, the differences in color makes for a greater separation of the scatter. This picture suggests, but doesn't conclusively establish, that there isn't much dependence between temperature and heart rate, and any dependence between temperature and heart rate isn't affected by gender. The scatter plot of the normalized data, in standard coordinates, on the *right* supports this view

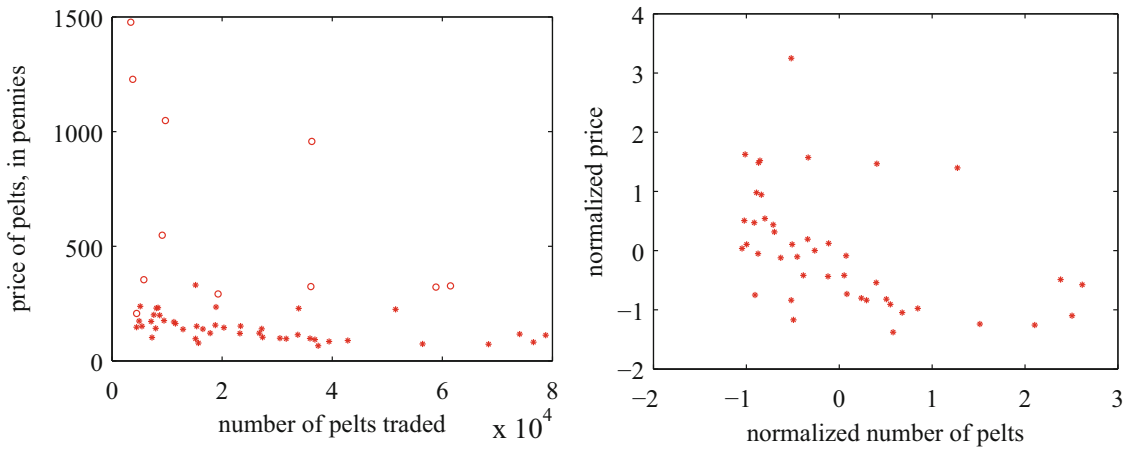

Fig. 2.13 *Left:* A scatter plot of the price of lynx pelts against the number of pelts (this is a repeat of Fig. 2.10 for reference). I have plotted data for 1901 to the end of the series as circles, and the rest of the data as *'s. It is quite hard to draw any conclusion from this data, because the scale is confusing. *Right:* A scatter plot of the price of pelts against the number of pelts for lynx pelts. I excluded data for 1901 to the end of the series, and then normalized both price and number of pelts. Notice that there is now a distinct trend; when there are fewer pelts, they are more expensive, and when there are more, they are cheaper

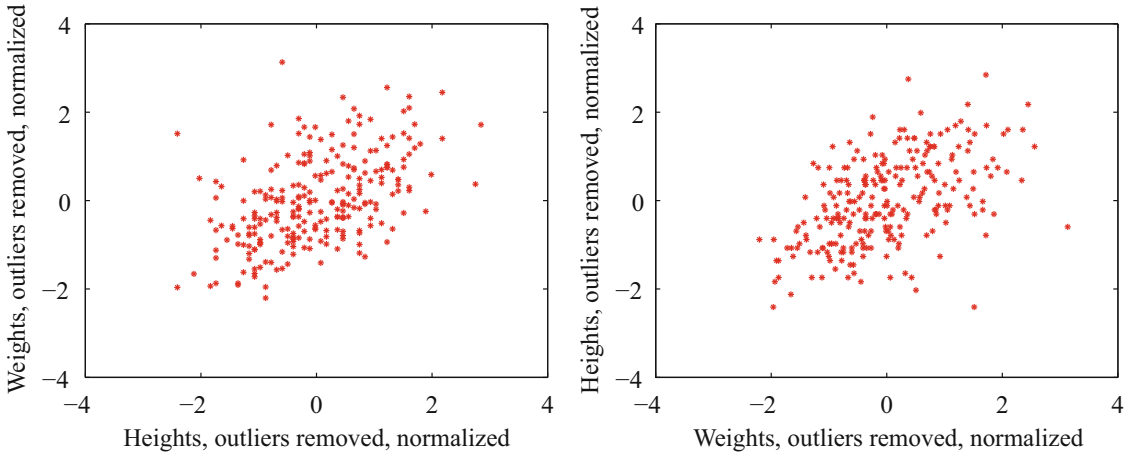

Fig. 2.14 On the *left*, a normalized scatter plot of weight (y-coordinate) against height (x-coordinate). On the *right*, a scatter plot of height (y-coordinate) against weight (x-coordinate). I've put these plots next to one another so you don't have to mentally rotate (which is what you should usually do)

or by imagining the new picture. The left plot tells you that data points with higher height value tend to have higher weight value; the right plot tells you that data points with higher weight value tend to have higher height value—i.e. the plots tell you the same thing. It doesn't really matter which one you look at. Again, the important word is "tend"—the plot doesn't tell you anything about *why*, it just tells you that when one variable is larger the other tends to be, too.

Positive correlation occurs when larger $\hat{x}$ values tend to appear with larger $\hat{y}$ values. This means that data points with small (i.e. negative with large magnitude) $\hat{x}$ values must have small $\hat{y}$ values, otherwise the mean of $\hat{x}$ (resp. $\hat{y}$) would be too big. In turn, this means that the scatter plot should look like a "smear" of data from the bottom left of the graph to the top right. The smear might be broad or narrow, depending on some details we'll discuss below. Figure 2.11 shows normalized scatter plots of weight against height, and of body temperature against heart rate. In the weight-height plot, you can clearly see that individuals who are higher tend to weigh more. The important word here is "tend"—taller people could be lighter, but mostly they tend not to be. Notice, also, that I did NOT say that they weighed more *because* they were taller, but only that they tend to be heavier.

Negative correlation occurs when larger $\hat{x}$ values tend to appear with smaller $\hat{y}$ values. This means that data points with small $\hat{x}$ values must have large $\hat{y}$ values, otherwise the mean of $\hat{x}$ (resp. $\hat{y}$) would be too big. In turn, this means that the scatter plot should look like a "smear" of data from the top left of the graph to the bottom right. The smear might be broad or narrow, depending on some details we'll discuss below. Figure 2.13 shows a normalized scatter plot of the lynx pelt-price data, where I have excluded the data from 1901 on. I did so because there seemed to be some other effect operating to drive prices up, which was inconsistent with the rest of the series. This plot suggests that when there were more pelts, prices were lower, as one would expect.

Zero correlation occurs when there is no relationship. This produces a characteristic shape in a scatter plot, but it takes a moment to understand why. If there really is no relationship, then knowing $\hat{x}$ will tell you nothing about $\hat{y}$. All we know is that $\mathsf{mean}\,(\{\hat{y}\}) = 0$, and $\mathsf{var}\,(\{\hat{y}\}) = 1$. This is enough information to predict what the plot will look like. We know that $\mathsf{mean}\,(\{\hat{x}\}) = 0$ and $\mathsf{var}\,(\{\hat{x}\}) = 1$; so there will be many data points with $\hat{x}$ value close to zero, and few with a much larger or much smaller $\hat{x}$ value. The same applies to $\hat{y}$. Now consider the data points in a strip of $\hat{x}$ values. If this strip is far away from the origin, there will be few data points in the strip, because there aren't many big $\hat{x}$ values. If there is no relationship, we don't expect to see large or small $\hat{y}$ values in this strip, because there are few data points in the strip and because large or small $\hat{y}$ values are uncommon—we see them only if there are many data points, and then seldom. So for a strip with $\hat{x}$ close to zero, we might see some $\hat{y}$ values that are far from zero because we will see many $\hat{y}$ values. For a strip with $\hat{x}$ that is far from zero, we expect to see few $\hat{y}$ values that are far from zero, because we see few points in this strip. This reasoning means the data should form a round blob, centered at the origin. In the temperature-heart rate plot of Fig. 2.12, it looks as though nothing of much significance is happening. The average heart rate seems to be about the same for people who run warm or who run cool. There is probably not much relationship here.

I have shown the three cases together in one figure using a real data example (Fig. 2.15), so you can compare the appearance of the plots.

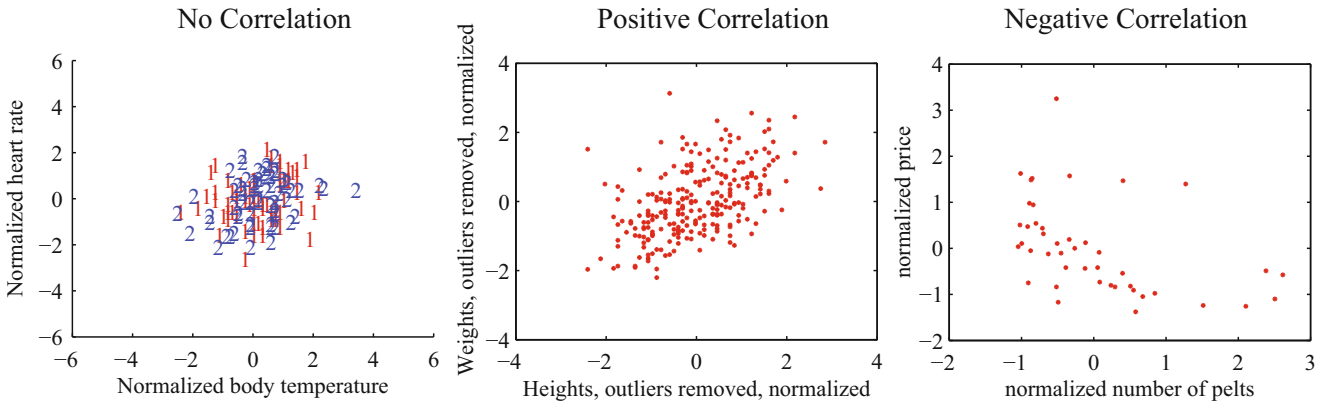

Fig. 2.15 The three kinds of scatter plot: I used the body temperature vs heart rate data for the zero correlation; the height-weight data for positive correlation; and the lynx data for negative correlation. The pictures aren't idealized—real data tends to be messy—but you can still see the basic structures

2.2.1 The Correlation Coefficient

Consider a normalized data set of N two-dimensional vectors. We can write the i'th data point *in standard coordinates* $(\hat{x}_i, \hat{y}_i)$. We already know many important summaries of this data, because it is in standard coordinates. We have $\mathsf{mean}\,(\{\hat{x}\}) = 0$; $\mathsf{mean}\,(\{\hat{y}\}) = 0$; $\mathsf{std}\,(\{\hat{x}\}) = 1$; and $\mathsf{std}\,(\{\hat{y}\}) = 1$. Each of these summaries is itself the mean of some monomial. So $\mathsf{std}\,(\{\hat{x}\})^2 = \mathsf{mean}\left(\{\hat{x}^2\}\right) = 1$; $\mathsf{std}\,(\{\hat{y}\})^2 = \mathsf{mean}\left(\{\hat{y}^2\}\right)$ (the other two are easy). We can rewrite this information in terms of means of monomials, giving $\mathsf{mean}\,(\{\hat{x}\}) = 0$; $\mathsf{mean}\,(\{\hat{y}\}) = 0$; $\mathsf{mean}\left(\{\hat{x}^2\}\right) = 1$; and $\mathsf{mean}\left(\{\hat{y}^2\}\right) = 1$. There is one monomial missing here, which is $\hat{x}\hat{y}$. The term $\mathsf{mean}\,(\{\hat{x}\hat{y}\})$ captures correlation between x and y. The term is known as the **correlation coefficient** or **correlation**.

Definition 2.1 (Correlation Coefficient) Assume we have N data items which are 2-vectors $(x_1, y_1), \ldots, (x_N, y_N)$, where $N > 1$. These could be obtained, for example, by extracting components from larger vectors. We compute the correlation coefficient by first normalizing the x and y coordinates to obtain $\hat{x}_i = \dfrac{(x_i - \mathsf{mean}\,(\{x\}))}{\mathsf{std}\,(x)}$, $\hat{y}_i = \dfrac{(y_i - \mathsf{mean}\,(\{y\}))}{\mathsf{std}\,(y)}$. The correlation coefficient is the mean value of $\hat{x}\hat{y}$, and can be computed as:

$$\mathsf{corr}\,(\{(x, y)\}) = \frac{\sum_i \hat{x}_i \hat{y}_i}{N}$$

Correlation is a measure of our ability to predict one value from another. The correlation coefficient takes values between -1 and 1 (we'll prove this below). If the correlation coefficient is close to 1, then we are likely to predict very well. Small correlation coefficients (under about 0.5, say, but this rather depends on what you are trying to achieve) tend not to be all that interesting, because (as we shall see) they result in rather poor predictions.

Figure 2.16 gives a set of scatter plots of different real data sets with different correlation coefficients. These all come from data set of age-height-weight, which you can find at http://www2.stetson.edu/~jrasp/data.htm (look for bodyfat.xls). In each case, two outliers have been removed. Age and height are hardly correlated, as you can see from the figure. Younger people do tend to be slightly taller, and so the correlation coefficient is -0.25. You should interpret this as a small correlation. However, the variable called "adiposity" (which isn't defined, but is presumably some measure of the amount of fatty tissue) is quite strongly correlated with weight, with a correlation coefficient is 0.86. Average tissue density is quite strongly negatively correlated with adiposity, because muscle is much denser than fat, so these variables are negatively correlated—we expect high density to appear with low adiposity, and vice versa. The correlation coefficient is -0.86. Finally, density is very strongly correlated with body weight. The correlation coefficient is -0.98.

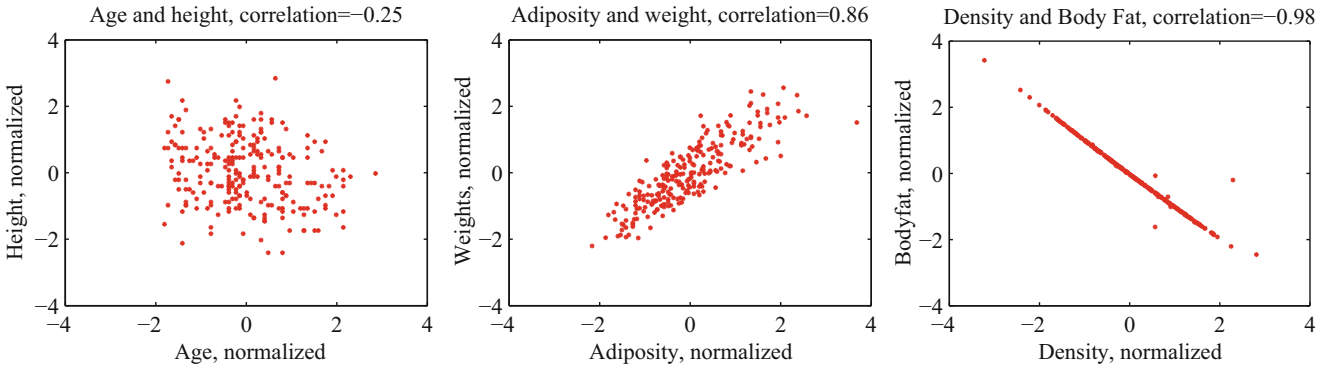

Fig. 2.16 Scatter plots for various pairs of variables for the age-height-weight dataset from http://www2.stetson.edu/~jrasp/data.htm; bodyfat.xls. In each case, two outliers have been removed, and the plots are in standard coordinates (compare to Fig. 2.17, which shows these data sets plotted in their original units). The legend names the variables

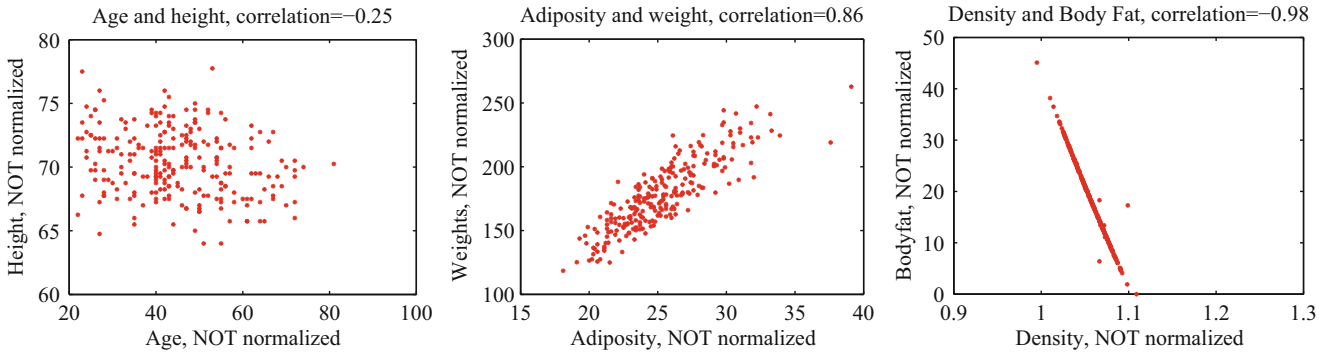

Fig. 2.17 Scatter plots for various pairs of variables for the age-height-weight dataset from http://www2.stetson.edu/~jrasp/data.htm; bodyfat.xls. In each case, two outliers have been removed, and the plots are *NOT* in standard coordinates (compare to Fig. 2.16, which shows these data sets plotted in normalized coordinates). The legend names the variables

It's not always convenient or a good idea to produce scatter plots in standard coordinates (among other things, doing so hides the units of the data, which can be a nuisance). Fortunately, scaling or translating data does not change the value of the correlation coefficient (though it can change the sign if one scale is negative). This means that it's worth being able to spot correlation in a scatter plot that isn't in standard coordinates (even though correlation is always *defined* in standard coordinates). Figure 2.17 shows different correlated datasets plotted in their original units. These data sets are the same as those used in Fig. 2.16.

You should memorize the properties of the correlation coefficient in the box. The first property is easy, and we relegate that to the exercises. One way to see that the correlation coefficient isn't changed by translation or scale is to notice that it is defined in standard coordinates, and scaling or translating data doesn't change those. Another way to see this is to scale and translate data, then write out the equations; notice that taking standard coordinates removes the effects of the scale and translation. In each case, notice that if the scale is negative, the sign of the correlation coefficient changes.

Useful Facts 2.1 (Properties of the Correlation Coefficient)

- The correlation coefficient is symmetric (it doesn't depend on the order of its arguments), so

$$\mathsf{corr}\left(\{(x,y)\}\right) = \mathsf{corr}\left(\{(y,x)\}\right)$$

(continued)

- The value of the correlation coefficient is not changed by translating the data. Scaling the data can change the sign, but not the absolute value. For constants $a \neq 0, b, c \neq 0, d$ we have

$$\mathsf{corr}\left(\{(ax+b, cx+d)\}\right) = \mathrm{sign}(ab)\mathsf{corr}\left(\{(x,y)\}\right)$$

- If $\hat{y}$ tends to be large (resp. small) for large (resp. small) values of $\hat{x}$, then the correlation coefficient will be positive.
- If $\hat{y}$ tends to be small (resp. large) for large (resp. small) values of $\hat{x}$, then the correlation coefficient will be negative.
- If $\hat{y}$ doesn't depend on $\hat{x}$, then the correlation coefficient is zero (or close to zero).
- The largest possible value is 1, which happens when $\hat{x} = \hat{y}$.
- The smallest possible value is -1, which happens when $\hat{x} = -\hat{y}$.

The property that, if $\hat{y}$ tends to be large (resp. small) for large (resp. small) values of $\hat{x}$, then the correlation coefficient will be positive, doesn't really admit a formal statement. But it's relatively straightforward to see what's going on. Because $\mathsf{mean}\left(\{\hat{x}\}\right) = 0$, small values of $\mathsf{mean}\left(\{\hat{x}\}\right)$ must be negative and large values must be positive. But $\mathsf{corr}\left(\{(x,y)\}\right) = \frac{\sum_i \hat{x}_i \hat{y}_i}{N}$; and for this sum to be positive, it should contain mostly positive terms. It can contain few or no hugely positive (or hugely negative) terms, because $\mathsf{std}\left(\hat{x}\right) = \mathsf{std}\left(\hat{y}\right) = 1$ so there aren't many large (or small) numbers. For the sum to contain mostly positive terms, then the sign of $\hat{x}_i$ should be the same as the sign $\hat{y}_i$ for most data items. Small changes to this argument work to show that if if $\hat{y}$ tends to be small (resp. large) for large (resp. small) values of $\hat{x}$, then the correlation coefficient will be negative.

Showing that no relationship means zero correlation requires slightly more work. Divide the scatter plot of the dataset up into thin vertical strips. There are S strips. Each strip is narrow, so the $\hat{x}$ value does not change much for the data points in a particular strip. For the s'th strip, write $N(s)$ for the number of data points in the strip, $\hat{x}(s)$ for the $\hat{x}$ value at the center of the strip, and $\bar{\hat{y}}(s)$ for the mean of the $\hat{y}$ values within that strip. Now the strips are narrow, so we can approximate all data points within a strip as having the same value of $\hat{x}$. This yields

$$\mathsf{mean}\left(\{\hat{x}\hat{y}\}\right) \approx \frac{1}{S} \sum_{s \in \text{strips}} \hat{x}(s)\left[N(s)\bar{\hat{y}}(s)\right]$$

(where you could replace $\approx$ with $=$ if the strips were narrow enough). Now assume that $\hat{y}(s)$ does not change from strip to strip, meaning that there is no relationship between $\hat{x}$ and $\hat{y}$ in this dataset (so the picture is like the left hand side in Fig. 2.15). Then each value of $\bar{\hat{y}}(s)$ is the same—we write $\bar{\hat{y}}$—and we can rearrange to get

$$\mathsf{mean}\left(\{\hat{x}\hat{y}\}\right) \approx \bar{\hat{y}} \frac{1}{S} \sum_{s \in \text{strips}} \hat{x}(s).$$

Now notice that

$$0 = \mathsf{mean}\left(\{\hat{y}\}\right) \approx \frac{1}{S} \sum_{s \in \text{strips}} N(s)\bar{\hat{y}}(s)$$

(where again you could replace $\approx$ with $=$ if the strips were narrow enough). This means that if every strip has the same value of $\bar{\hat{y}}(s)$, then that value must be zero. In turn, if there is no relationship between $\hat{x}$ and $\hat{y}$, we must have $\mathsf{mean}\left(\{\hat{x}\hat{y}\}\right) = 0$.

Property 2.1 The largest possible value of the correlation is 1, and this occurs when $\hat{x}_i = \hat{y}_i$ for all i. The smallest possible value of the correlation is -1, and this occurs when $\hat{x}_i = -\hat{y}_i$ for all i.

Proposition

$$-1 \leq \mathsf{corr}\left(\{(x,y)\}\right) \leq 1$$

(continued)

Proof Writing $\hat{x}, \hat{y}$ for the normalized coefficients, we have

$$\mathsf{corr}\left(\{(x,y)\}\right) = \frac{\sum_i \hat{x}_i \hat{y}_i}{N}$$

and you can think of the value as the inner product of two vectors. We write

$$\mathbf{x} = \frac{1}{\sqrt{N}}\left[\hat{x}_1, \hat{x}_2, \ldots \hat{x}_N\right] \text{ and}$$
$$\mathbf{y} = \frac{1}{\sqrt{N}}\left[\hat{y}_1, \hat{y}_2, \ldots \hat{y}_N\right]$$

and we have $\mathsf{corr}\left(\{(x,y)\}\right) = \mathbf{x}^T\mathbf{y}$. Notice $\mathbf{x}^T\mathbf{x} = \mathsf{std}\left(x\right)^2 = 1$, and similarly for $\mathbf{y}$. But the inner product of two vectors is at its maximum when the two vectors are the same, and this maximum is 1. This argument is also sufficient to show that smallest possible value of the correlation is -1, and this occurs when $\hat{x}_i = -\hat{y}_i$ for all i.

2.2.2 Using Correlation to Predict

Assume we have N data items which are 2-vectors $(x_1, y_1), \ldots, (x_N, y_N)$, where $N > 1$. These could be obtained, for example, by extracting components from larger vectors. As usual, we will write $\hat{x}_i$ for x_i in normalized coordinates, and so on. Now assume that we know the correlation coefficient is r (this is an important, traditional notation). What does this mean?

One (very useful) interpretation is in terms of prediction. Assume we have a data point $(x_0, ?)$ where we know the x-coordinate, but not the y-coordinate. We can use the correlation coefficient to predict the y-coordinate. First, we transform to standard coordinates. Now we must obtain the best $\hat{y}_0$ value to predict, using the $\hat{x}_0$ value we have.

We want to construct a prediction function which gives a prediction for any value of $\hat{x}$. This predictor should behave as well as possible on our existing data. For each of the $(\hat{x}_i, \hat{y}_i)$ pairs in our data set, the predictor should take $\hat{x}_i$ and produce a result as close to $\hat{y}_i$ as possible. We can choose the predictor by looking at the errors it makes at each data point.

We write $\hat{y}_i^p$ for the value of $\hat{y}_i$ predicted at $\hat{x}_i$. The simplest form of predictor is linear. If we predict using a linear function, then we have, for some unknown a, b, that $\hat{y}_i^p = a\hat{x}_i + b$. Now think about $u_i = \hat{y}_i - \hat{y}_i^p$, which is the error in our prediction. We would like to have $\mathsf{mean}\left(\{u\}\right) = 0$ (otherwise, we could reduce the error of the prediction just by subtracting a constant).

$$\begin{aligned}
\mathsf{mean}\left(\{u\}\right) &= \mathsf{mean}\left(\{\hat{y} - \hat{y}^p\}\right) \\
&= \mathsf{mean}\left(\{\hat{y}\}\right) - \mathsf{mean}\left(\{a\hat{x}_i + b\}\right) \\
&= \mathsf{mean}\left(\{\hat{y}\}\right) - a\mathsf{mean}\left(\{\hat{x}\}\right) + b \\
&= 0 - a0 + b \\
&= 0.
\end{aligned}$$

This means that we must have $b = 0$.

To estimate a, we need to think about $\mathsf{var}\left(\{u\}\right)$. We should like $\mathsf{var}\left(\{u\}\right)$ to be as small as possible, so that the errors are as close to zero as possible (remember, small variance means small standard deviation which means the data is close to the mean). We have

$$\begin{aligned}
\mathsf{var}\,(\{u\}) &= \mathsf{var}\,(\{\hat{y} - \hat{y}^p\}) \\
&= \mathsf{mean}\left(\left\{(\hat{y} - a\hat{x})^2\right\}\right) \quad \text{because } \mathsf{mean}\,(\{u\}) = 0 \\
&= \mathsf{mean}\left(\left\{(\hat{y})^2 - 2a\hat{x}\hat{y} + a^2(\hat{x})^2\right\}\right) \\
&= \mathsf{mean}\left(\left\{(\hat{y})^2\right\}\right) - 2a\mathsf{mean}\,(\{\hat{x}\hat{y}\}) + a^2\mathsf{mean}\left(\left\{(\hat{x})^2\right\}\right) \\
&= 1 - 2ar + a^2,
\end{aligned}$$

which we want to minimize by choice of a. At the minimum, we must have

$$\frac{d\mathsf{var}\,(\{u_i\})}{da} = 0 = -2r + 2a$$

so that $a = r$ and the correct prediction is

$$\hat{y}_0^p = r\hat{x}_0$$

You can use a version of this argument to establish that if we have $(?, \hat{y}_0)$, then the best prediction for $\hat{x}_0$ (*which is in standard coordinates*) is $r\hat{y}_0$. It is important to notice that the coefficient of $\hat{y}_0$ is NOT $1/r$; you should work this example, which appears in the exercises. We now have a prediction procedure, outlined below.

Procedure 2.1 (Predicting a Value Using Correlation) Assume we have N data items which are 2-vectors $(x_1, y_1), \ldots, (x_N, y_N)$, where $N > 1$. These could be obtained, for example, by extracting components from larger vectors. Assume we have an x value x_0 for which we want to give the best prediction of a y value, based on this data. The following procedure will produce a prediction:

- Transform the data set into standard coordinates, to get

$$\begin{aligned}
\hat{x}_i &= \frac{1}{\mathsf{std}\,(x)}(x_i - \mathsf{mean}\,(\{x\})) \\
\hat{y}_i &= \frac{1}{\mathsf{std}\,(y)}(y_i - \mathsf{mean}\,(\{y\})) \\
\hat{x}_0 &= \frac{1}{\mathsf{std}\,(x)}(x_0 - \mathsf{mean}\,(\{x\})).
\end{aligned}$$

- Compute the correlation

$$r = \mathsf{corr}\,(\{(x, y)\}) = \mathsf{mean}\,(\{\hat{x}\hat{y}\}).$$

- Predict $\hat{y}_0 = r\hat{x}_0$.
- Transform this prediction into the original coordinate system, to get

$$y_0 = \mathsf{std}\,(y)r\hat{x}_0 + \mathsf{mean}\,(\{y\})$$

Now assume we have a y value y_0, for which we want to give the best prediction of an x value, based on this data. The following procedure will produce a prediction:

- Transform the data set into standard coordinates.
- Compute the correlation.
- Predict $\hat{x}_0 = r\hat{y}_0$.
- Transform this prediction into the original coordinate system, to get

$$x_0 = \mathsf{std}\,(x)r\hat{y}_0 + \mathsf{mean}\,(\{x\})$$

There is another way of thinking about this prediction procedure, which is often helpful. Assume we need to predict a value for x_0. In normalized coordinates, our prediction is $\hat{y}^p = r\hat{x}_0$; if we revert back to the original coordinate system, the prediction becomes

$$\frac{(y^p - \text{mean}(\{y\}))}{\text{std}(y)} = r\left(\frac{(x_0 - \text{mean}(\{x\}))}{\text{std}(x)}\right).$$

This gives a really useful rule of thumb, which I have broken out in the box below.

Procedure 2.2 (Predicting a Value Using Correlation: Rule of Thumb—1) If x_0 is k standard deviations from the mean of x, then the predicted value of y will be rk standard deviations away from the mean of y, and the sign of r tells whether y increases or decreases.

An even more compact version of the rule of thumb is in the following box.

Procedure 2.3 (Predicting a Value Using Correlation: Rule of Thumb—2) The predicted value of y goes up by r standard deviations when the value of x goes up by one standard deviation.

We can compute the average root mean square error that this prediction procedure will make. The square of this error must be

$$\begin{aligned}\text{mean}\left(\{u^2\}\right) &= \text{mean}\left(\{y^2\}\right) - 2r\text{mean}\left(\{xy\}\right) + r^2\text{mean}\left(\{x^2\}\right)\\ &= 1 - 2r^2 + r^2\\ &= 1 - r^2\end{aligned}$$

so the root mean square error will be $\sqrt{1-r^2}$. This is yet another interpretation of correlation; if x and y have correlation close to one, then predictions could have very small root mean square error, and so might be very accurate. In this case, knowing one variable is about as good as knowing the other. If they have correlation close to zero, then the root mean square error in a prediction might be as large as the root mean square error in $\hat{y}$—which means the prediction is nearly a pure guess.

The prediction argument means that we can spot correlations for data in other kinds of plots—one doesn't have to make a scatter plot. For example, if we were to observe a child's height from birth to their 10'th year (you can often find these observations in ballpen strokes, on kitchen walls), we could plot height as a function of year. If we also had their weight (less easily found), we could plot weight as a function of year, too. The prediction argument above say that, if you can predict the weight from the height (or vice versa) then they're correlated. One way to spot this is to look and see if one curve goes up when the other does (or goes down when the other goes up). You can see this effect in Fig. 2.5, where (before 19h00), prices go down when the number of pelts goes up, and vice versa. These two variables are negatively correlated.

2.2.3 Confusion Caused by Correlation

There is one very rich source of potential (often hilarious) mistakes in correlation. When two variables are correlated, they change together. If the correlation is positive, that means that, in typical data, if one is large then the other is large, and if one is small the other is small. In turn, this means that one can make a reasonable prediction of one from the other. However, correlation DOES NOT mean that changing one variable causes the other to change (sometimes known as causation).

Two variables in a dataset could be correlated for a variety of reasons. One important reason is pure accident. If you look at enough pairs of variables, you may well find a pair that appears to be correlated just because you have a small set of observations. Imagine, for example, you have a dataset consisting of only two high dimensional vectors—there is a pretty good chance that there is some correlation between the components. Such accidents can occur in large datasets, particularly if the dimensions are high.

Another reason variables could be correlated is that there is some causal relationship—for example, pressing the accelerator tends to make the car go faster, and so there will be some correlation between accelerator position and car

acceleration. As another example, adding fertilizer does tend to make a plant grow bigger. Imagine you record the amount of fertilizer you add to each pot, and the size of the resulting potplant. There should be some correlation.

Yet another reason variables could be correlated is that there is some other background variable—often called a **latent variable**—linked causally to each of the observed variables. For example, in children (as Freedman, Pisani and Purves note in their excellent *Statistics*), shoe size is correlated with reading skills. This DOES NOT mean that making your feet grow will make you read faster, or that you can make your feet shrink by forgetting how to read. The real issue here is the age of the child. Young children tend to have small feet, and tend to have weaker reading skills (because they've had less practice). Older children tend to have larger feet, and tend to have stronger reading skills (because they've had more practice). You can make a reasonable prediction of reading skills from foot size, because they're correlated, even though there is no direct connection.

This kind of effect can mask correlations, too. Imagine you want to study the effect of fertilizer on potplants. You collect a set of pots, put one plant in each, and add different amounts of fertilizer. After some time, you record the size of each plant. You expect to see correlation between fertilizer amount and plant size. But you might not if you had used a different species of plant in each pot. Different species of plant can react quite differently to the same fertilizer (some plants just die if over-fertilized), so the species could act as a latent variable. With an unlucky choice of the different species, you might even conclude that there was a negative correlation between fertilizer and plant size. This sort of thing happens often, and it's an effect you should watch out for.

2.3 Sterile Males in Wild Horse Herds

Large herds of wild horses are (apparently) a nuisance, but keeping down numbers by simply shooting surplus animals would provoke outrage. One strategy that has been adopted is to sterilize males in the herd; if a herd contains sufficient sterile males, fewer foals should result. But catching stallions, sterilizing them, and reinserting them into a herd is a performance—does this strategy work?

We can get some insight by plotting data. At http://lib.stat.cmu.edu/DASL/Datafiles/WildHorses.html, you can find a dataset covering herd management in wild horses. I have plotted part of this dataset in Fig. 2.18. In this dataset, there are counts of all horses, sterile males, and foals made on each of a small number of days in 1986, 1987, and 1988 for each of two herds. I extracted data for one herd. I have plotted this data as a function of the count of days since the first data point, because this makes it clear that some measurements were taken at about the same time, but there are big gaps in the measurements. In this plot, the data points are shown with a marker. Joining them leads to a confusing plot because the data points vary quite strongly. However, notice that the size of the herd drifts down slowly (you could hold a ruler against the plot to see the trend), as does the number of foals, when there is a (roughly) constant number of sterile males.

Does sterilizing males result in fewer foals? This is likely hard to answer for this dataset, but we could ask whether herds with more sterile males have fewer foals. A scatter plot is a natural tool to attack this question. However, the scatter plots of Fig. 2.19 suggest, rather surprisingly, that when there are more sterile males there are more adults (and vice versa), and when there are more sterile males there are more foals (and vice versa). This is borne out by a correlation analysis. The correlation

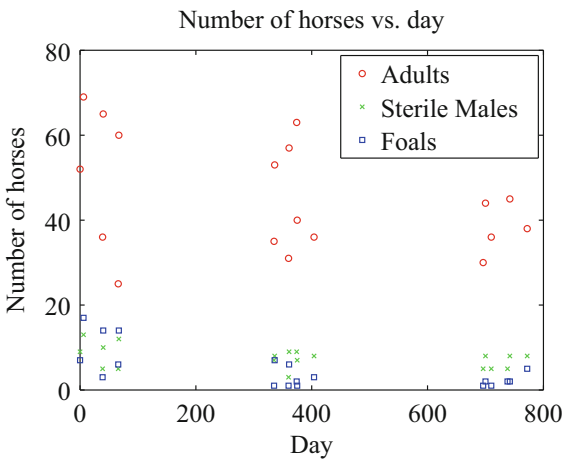

Fig. 2.18 A plot of the number of adult horses, sterile males, and foals in horse herds over a period of 3 years. The plot suggests that introducing sterile males might cause the number of foals to go down. Data from http://lib.stat.cmu.edu/DASL/Datafiles/WildHorses.html

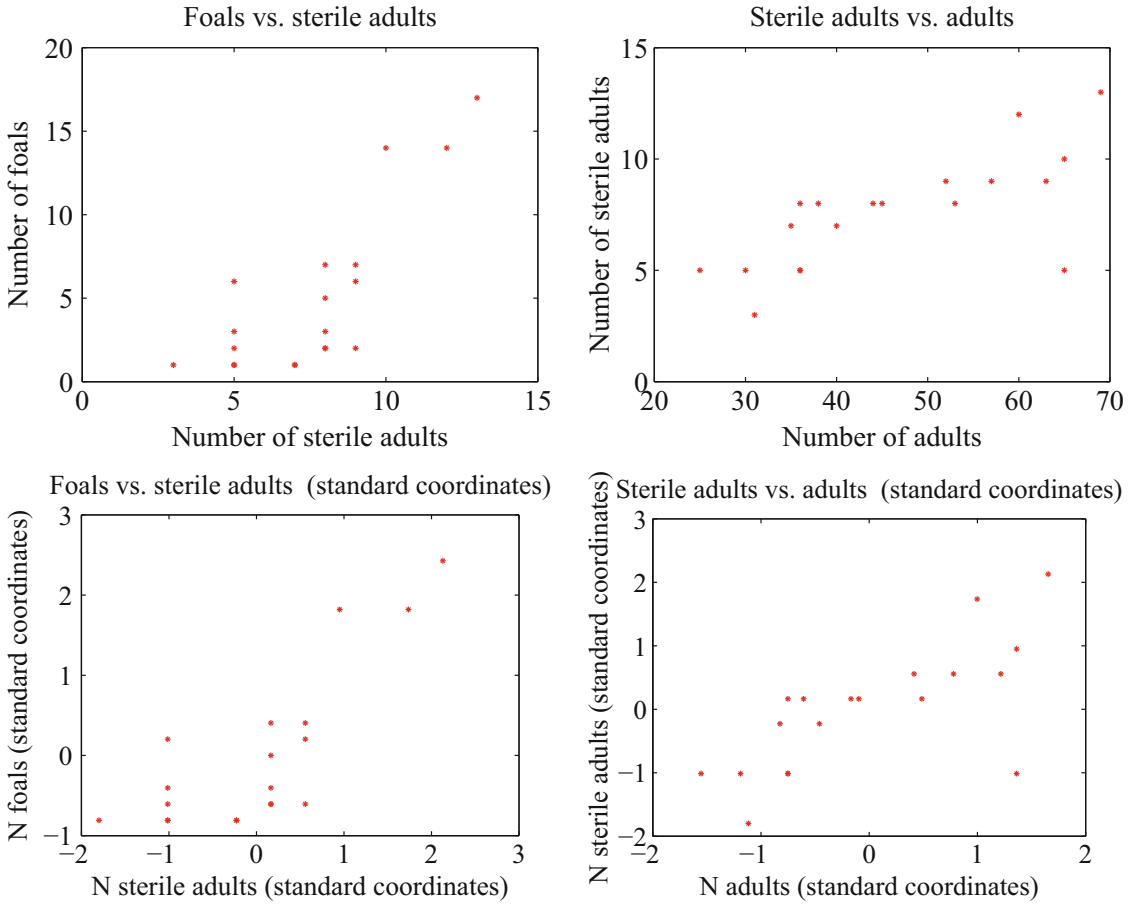

Fig. 2.19 Scatter plots of the number of sterile males in a horse herd against the number of adults, and the number of foals against the number of sterile males, from data of http://lib.stat.cmu.edu/DASL/Datafiles/WildHorses.html. *Top:* unnormalized; *bottom:* standard coordinates

coefficient between foals and sterile males is 0.74, and the correlation coefficient between adults and sterile males is 0.68. You should find this very surprising—how do the horses know how many sterile males there are in the herd? You might think that this is an effect of scaling the plot, but there is a scatter plot in normalized coordinates in Fig. 2.19 that is entirely consistent with the conclusions suggested by the unnormalized plot. What is going on here?

The answer is revealed by the scatter plots of Fig. 2.20. Here, rather than plotting a '*' at each data point, I have plotted the day number of the observation. This is in days from the first observation. You can see that the whole herd is shrinking—observations where there are many adults (resp. sterile adults, foals) occur with small day numbers, and observations where there are few have large day numbers. Because the whole herd is shrinking, it is true that when there are more adults and more sterile males, there are also more foals. Alternatively, you can see the plots of Fig. 2.18 as a scatter plot of herd size (resp. number of foals, number of sterile males) against day number. Then it becomes clear that the whole herd is shrinking, as is the size of each group. To drive this point home, we can look at the correlation coefficient between adults and days (−0.24), between sterile adults and days (−0.37), and between foals and days (−0.61). We can use the rule of thumb in box 2.3 to interpret this. This means that every 282 days, the herd loses about three adults; about one sterile adult; and about three foals. For the herd to have a stable size, it needs to gain by birth as many foals as it loses both to growing up and to death. If the herd is losing three foals every 282 days, then if they all grow up to replace the missing adults, the herd will be shrinking slightly (because it is losing four adults in this time); but if it loses foals to natural accidents, etc., then it is shrinking rather fast.

The message of this example is important. To understand a simple dataset, you might need to plot it several ways. You should make a plot, look at it and ask what it says, and then try to use another type of plot to confirm or refute what you think might be going on.

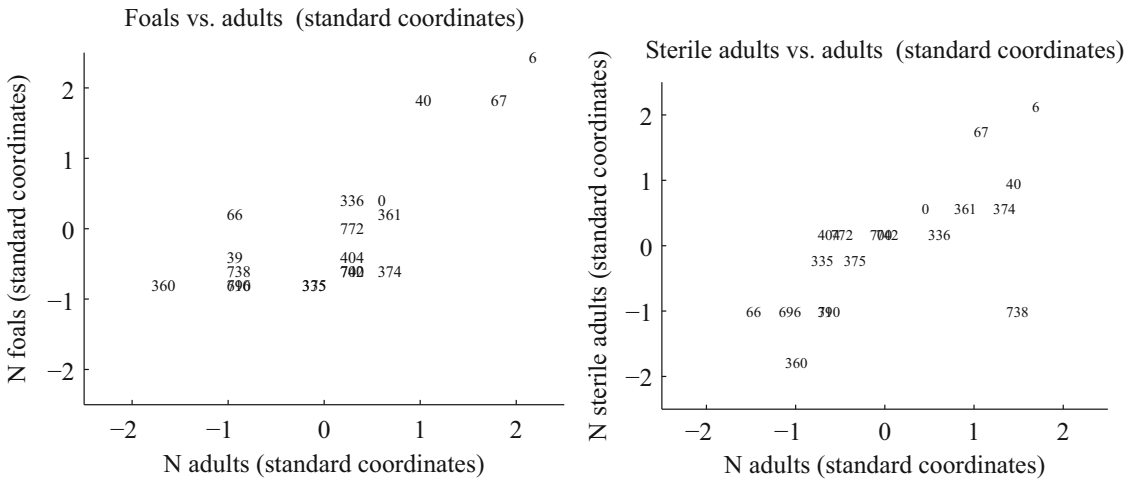

Fig. 2.20 Scatter plots of the number of foals vs. the number of adults and the number of adults vs. the number of sterile adults for the wild horse herd, from http://lib.stat.cmu.edu/DASL/Datafiles/WildHorses.html. Rather than plot data points as dots, I have plotted the *day* on which the observation was made. Notice how the herd starts large, and then shrinks

2.4 You Should

2.4.1 Remember These Definitions

2.4.2 Remember These Terms

2.4.3 Remember These Facts

2.4.4 Use These Procedures

2.4.5 Be Able to

- Plot a bar chart, a heat map, and a pie chart for a categorical dataset.
- Plot a dataset as a graph, making sensible choices about markers, lines and the like.

- Plot a scatter plot for a dataset.
- Plot a normalized scatter plot for a dataset.
- Interpret the scatter plot to tell the sign of the correlation between two variables, and estimate the size of the correlation coefficient.
- Compute a correlation coefficient.
- Interpret a correlation coefficient.
- Use correlation to make predictions.

Problems

2.1 In a population, the correlation coefficient between weight and adiposity is 0.9. The mean weight is 150lb. The standard deviation in weight is 30lb. Adiposity is measured on a scale such that the mean is 0.8, and the standard deviation is 0.1.

(a) Using this information, predict the expected adiposity of a subject whose weight is 170lb
(b) Using this information, predict the expected weight of a subject whose adiposity is 0.75
(c) How reliable do you expect this prediction to be? Why? (your answer should be a property of correlation, not an opinion about adiposity or weight)

2.2 In a population, the correlation coefficient between family income and child IQ is 0.30. The mean family income was \$60,000. The standard deviation in income is \$20,000. IQ is measured on a scale such that the mean is 100, and the standard deviation is 15.

(a) Using this information, predict the expected IQ of a child whose family income is \$70,000
(b) How reliable do you expect this prediction to be? Why? (your answer should be a property of correlation, not an opinion about IQ)
(c) The family income now rises—does the correlation predict that the child will have a higher IQ? Why?

2.3 Show that $\mathsf{corr}(\{(x, y)\}) = \mathsf{corr}(\{(y, x)\})$ by substituting into the definition.

2.4 Show that if $\hat{y}$ tends to be small (resp. large) for large (resp. small) values of $\hat{x}$, then the correlation coefficient will be negative.

2.5 We have a 2D dataset consisting of N pairs $(\hat{x}_i, \hat{y}_i)$ in normalized coordinates. This data has correlation coefficient r. We observe a new $\hat{y}$ value $\hat{y}_0$, and wish to predict the (unknown) x value. We will do so with a linear prediction, choosing a, b, to predict an $\hat{x}$ for any $\hat{y}$ using the rule $\hat{x}^p = a\hat{y}^p + b$. Write $u_i = \hat{x}_i - \hat{x}_i^p$ for the error that this rule makes on each data item.

(a) We require $\mathsf{mean}(\{u\}) = 0$. Show that this means that $b = 0$.
(b) We require that $\mathsf{var}(\{u\})$ is minimized. Show that this means that $a = r$.
(c) We now have a result that seems paradoxical—if I have $(\hat{x}_0, ?)$ I predict $(\hat{x}_0, r\hat{x}_0)$ and if I have $(?, y_0)$, I predict $(r\hat{y}_0, \hat{y}_0)$. Use Fig. 2.21 to explain why this is right. The important difference between the two lines is that lies (approximately) in the middle of each vertical span of data, and the other lies (approximately) in the middle of each horizontal span of data.

2.6 I did the programming exercise about the earth temperature below. I looked at the years 1965–2012. Write $\{(y, T)\}$ for the dataset giving the temperature (T) of the earth in year y. I computed: $\mathsf{mean}(\{y\}) = 1988.5$, $\mathsf{std}(y) = 14$, $\mathsf{mean}(\{T\}) = 0.175$, $\mathsf{std}(T) = 0.231$ and $\mathsf{corr}(\{y\})T = 0.892$. What is the best prediction using this information for the temperature in mid 2014? in mid 2028? in mid 2042?

2.7 I did the programming exercise about the earth temperature below. It is straightforward to build a dataset $\{(T, n_t)\}$ where each entry contains the temperature of the earth (T) and the number of counties where FEMA declared tornadoes n_t (for each year, you look up T and n_t, and make a data item). I computed: $\mathsf{mean}(\{T\}) = 0.175$, $\mathsf{std}(T) = 0.231$, $\mathsf{mean}(\{n_t\}) = 31.6$,

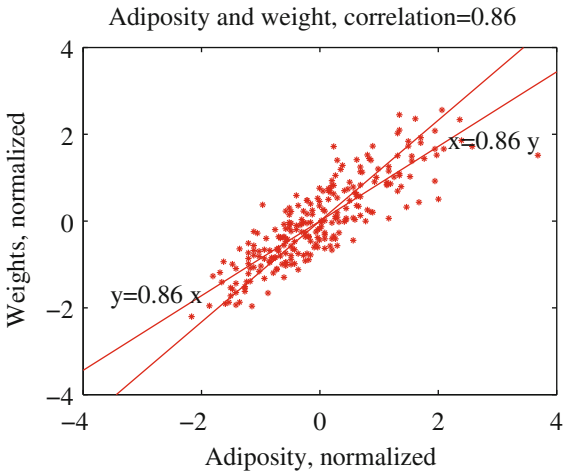

Fig. 2.21 This figure shows two lines, $y = 0.86x$ and $x = 0.86y$, superimposed on the normalized adiposity-weight scatter plot

$\mathsf{std}(n_t) = 30.8$, and $\mathsf{corr}(\{T\})n_t = 0.471$. What is the best prediction using this information for the number of tornadoes if the global earth temperature is 0.5? 0.6? 0.7?

Programming Exercises

2.8 At http://lib.stat.cmu.edu/DASL/Datafiles/cigcancerdat.html, you will find a dataset recording per capita cigarette sales and cancer deaths per 100 K population for a variety of cancers, recorded for 43 states and the District of Columbia in 1960.

(a) Plot a scatter plot of lung cancer deaths against cigarette sales, using the two letter abbreviation for each state as a marker. You should see two fairly obvious outliers. The backstory at http://lib.stat.cmu.edu/DASL/Stories/cigcancer.html suggests that the unusual sales in Nevada are generated by tourism (tourists go home, and die there) and the unusual sales in DC are generated by commuting workers (who also die at home).
(b) What is the correlation coefficient between per capita cigarette sales and lung cancer deaths per 100 K population? Compute this with, and without the outliers. What effect did the outliers have? Why?
(c) What is the correlation coefficient between per capita cigarette sales and bladder cancer deaths per 100 K population? Compute this with, and without the outliers. What effect did the outliers have? Why?
(d) What is the correlation coefficient between per capita cigarette sales and kidney cancer deaths per 100 K population? Compute this with, and without the outliers. What effect did the outliers have? Why?
(e) What is the correlation coefficient between per capita cigarette sales and leukemia deaths per 100 K population? Compute this with, and without the outliers. What effect did the outliers have? Why?
(f) You should have computed a positive correlation between cigarette sales and lung cancer deaths. Does this mean that smoking causes lung cancer? Why?
(g) You should have computed a negative correlation between cigarette sales and leukemia deaths. Does this mean that smoking cures leukemia? Why?

2.9 At http://www.cru.uea.ac.uk/cru/info/warming/gtc.csv, you can find a dataset of global temperature by year. When I accessed this, the years spanned 1880–2012. I don't know what units the temperatures are measured in. Keep in mind that measuring the temperature of the earth has non-trivial difficulties (you can't just insert an enormous thermometer!), and if you look at http://www.cru.uea.ac.uk/cru and http://www.cru.uea.ac.uk/cru/data/temperature/ you can see some discussion of the choices made to get these measurements. There are two kinds of data in this dataset, smoothed and unsmoothed. I used the unsmoothed data, which should be fine for our purposes. The government publishes a great deal of data at http://data.gov. From there, I found a dataset, published by the Federal Emergency Management Agency (FEMA), of all federally declared disasters (which I found at http://www.fema.gov/media-library/assets/documents/28318?id=6292). We would like to see whether weather related disasters are correlated to global temperature.

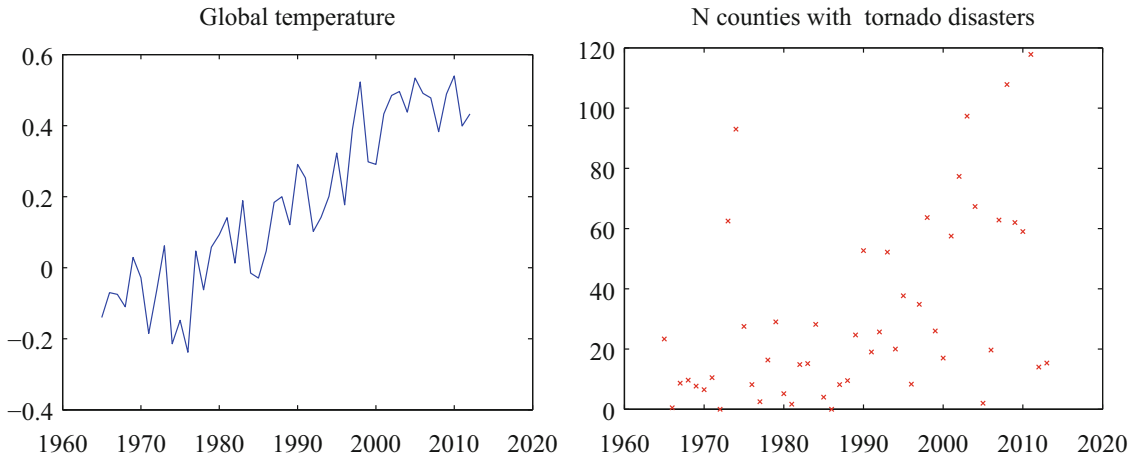

Fig. 2.22 Plots I prepared from *left* uea data on temperature and *right* FEMA data on tornadoes by county. These should help you tell if you're on the right track

(a) The first step is preprocessing the data. The FEMA data has all sorts of information. From 1965 on, a disaster was declared per county (so you get one line in the data set for each county), but before 1965, it seems to have been by state. We divide the disasters into four types: TORNADO, FLOOD, STORM, HURRICANE. (FEMA seems to have a much richer type system). We want to know how many counties declare a disaster of each type, in each year. This is a really rough estimate of the number of people affected by the disaster. If a disaster has two types (in some rows, you will see "SEVERE STORMS, HEAVY RAINS & FLOODING" we will allocate the credit evenly between the types (i.e. for this case we would count 1/2 for STORM and 1/2 for FLOOD). You should write code that will (a) read the dataset and then (b) compute a table with the count of the number of counties where a disaster of each type has occurred for each year. This takes a bit of work. Notice you only need to deal with two columns of the data (the date it was declared, and the type). Notice also that FEMA changed the way it represented dates somewhere through the first column (they added times), which can cause problems. You can tell the type of the disaster by just using a string match routine with the four keywords. Figure 2.22 shows the plot of temperature and of number of counties where FEMA declared a tornado disaster for this data.

(b) Plot a normalized scatter plot of the number of counties where FEMA declared the disaster against temperature, for each kind.

(c) For each kind of disaster, compute the correlation coefficient between the number of counties where FEMA declared the disaster and the year. For each kind of disaster, use this correlation coefficient to predict the number of disasters of this kind for 2013. Compare this to the true number, and explain what you see.

(d) For each kind of disaster, compute the correlation coefficient between the number of counties where FEMA declared the disaster and the global temperature. For each kind of disaster, use this correlation coefficient to predict the number of disasters of this kind when the earth reaches 0.6 temperature units and 0.7 temperature units (on the absolute temperature scale).

(e) Does this data show that warming of the earth causes weather disasters? Why?

(f) Does this data suggest that more people will be affected by disasters in the US in the future? Why?

(g) Does this data suggest that the earth will be warmer in the future? Why?

2.10 If you go to https://github.com/TheUpshot/Military-Surplus-Gear, you will find data on purchases of military weapons by US police departments. This data is organized by state and county. There's a fair amount of data here, and you'll need to do some data jockeying.

(a) Prepare a plot showing how much each Illinois county spent under this program.

(b) Now look up population levels in the counties. Prepare a plot showing how much each county spent *per capita*.

(c) Prepare a graphic illustrating what the overall most popular items were—i.e., those items counties bought the most of.

(d) Prepare a graphic illustrating on what items the most money was spent—for example, was more money spent on "RIFLE, 5.56 MILLIMETER" or on "MINE RESISTANT VEHICLE"?

(e) Prepare a graphic illustrating the pattern of purchases across counties for the ten overall most popular items.

(f) Can you draw any interesting conclusions?