Contents

CHAPTER 1

Notation and conventions

A dataset as a collection of d-tuples (a d-tuple is an ordered list of d elements). Tuples differ from vectors, because we can always add and subtract vectors, but we cannot necessarily add or subtract tuples. There are always N items in any dataset. There are always d elements in each tuple in a dataset. The number of elements will be the same for every tuple in any given tuple. Sometimes we may not know the value of some elements in some tuples.

We use the same notation for a tuple and for a vector. Most of our data will be vectors. We write a vector in bold, so $\mathbf{x}$ could represent a vector or a tuple (the context will make it obvious which is intended).

The entire data set is $\{\mathbf{x}\}$. When we need to refer to the i'th data item, we write $\mathbf{x}_i$. Assume we have N data items, and we wish to make a new dataset out of them; we write the dataset made out of these items as $\{\mathbf{x}_i\}$ (the i is to suggest you are taking a set of items and making a dataset out of them). If we need to refer to the j'th component of a vector $\mathbf{x}_i$, we will write $x_i^{(j)}$ (notice this isn't in bold, because it is a component not a vector, and the j is in parentheses because it isn't a power). Vectors are always column vectors.

Terms:

- $\mathsf{mean}\left(\{x\}\right)$ is the mean of the dataset $\{x\}$ (definition 1, page 20).
- $\mathsf{std}\left(x\right)$ is the standard deviation of the dataset $\{x\}$ (definition 2, page 22).
- $\mathsf{var}\left(\{x\}\right)$ is the standard deviation of the dataset $\{x\}$ (definition 3, page 25).
- $\mathsf{median}\left(\{x\}\right)$ is the standard deviation of the dataset $\{x\}$ (definition 4, page 26).
- $\mathsf{percentile}(\{x\}, k)$ is the $k\%$ percentile of the dataset $\{x\}$ (definition 5, page 27).
- $\mathsf{iqr}\{x\}$ is the interquartile range of the dataset $\{x\}$ (definition 7, page 27).
- $\{\hat{x}\}$ is the dataset $\{x\}$, transformed to standard coordinates (definition 8, page 32).
- Standard normal data is defined in definition 9, page 33).
- Normal data is defined in definition 10, page 34).
- $\mathsf{corr}\left(\{(x, y)\}\right)$ is the correlation between two components x and y of a dataset (definition 11, page 44).
- $\emptyset$ is the empty set.
- Ω is the set of all possible outcomes of an experiment.
- Sets are written as $\mathcal{A}$.

- $\mathcal{A}^c$ is the complement of the set $\mathcal{A}$ (i.e. $\Omega - \mathcal{A}$).
- $\mathcal{E}$ is an event (page 50).
- $P(\{\mathcal{E}\})$ is the probability of event $\mathcal{E}$ (page 50).
- $P(\{\mathcal{E}\}|\{\mathcal{F}\})$ is the probability of event $\mathcal{E}$, conditioned on event $\mathcal{F}$ (page 50).
- $p(x)$ is the probability that random variable X will take the value x; also written $P(\{X = x\})$ (page 50).
- $p(x, y)$ is the probability that random variable X will take the value x and random variable Y will take the value y; also written $P(\{X = x\} \cap \{Y = y\})$ (page 50).
- $\underset{x}{\text{argmax}} \, f(x)$ means the value of x that maximises $f(x)$.
- $\hat{\theta}$ is an estimated value of a parameter θ.

Background information:

- *Cards:* A standard deck of playing cards contains 52 cards. These cards are divided into four suits. The suits are: spades and clubs (which are black); and hearts and diamonds (which are red). Each suit contains 13 cards: Ace, 2, 3, 4, 5, 6, 7, 8, 9, 10, Jack (sometimes called Knave), Queen and King. It is common to call Jack, Queen and King *court cards*.
- *Dice:* If you look hard enough, you can obtain dice with many different numbers of sides (though I've never seen a three sided die). We adopt the convention that the sides of an N sided die are labeled with the numbers $1 \ldots N$, and that no number is used twice. Most dice are like this.
- *Fairness:* Each face of a fair coin or die has the same probability of landing upmost in a flip or roll.

1.1 SOME USEFUL MATHEMATICAL FACTS

The gamma function $\Gamma(x)$ is defined by a series of steps. First, we have that for n an integer,

$$\Gamma(n) = (n-1)!$$

and then for z a complex number with positive real part (which includes positive real numbers), we have

$$\Gamma(z) = \int_0^\infty t^z \frac{e^{-t}}{t} dt.$$

By doing this, we get a function on positive real numbers that is a smooth interpolate of the factorial function. We won't do any real work with this function, so won't expand on this definition. In practice, we'll either look up a value in tables or require a software environment to produce it.

1.2 ACKNOWLEDGEMENTS

Typos spotted by: Han Chen (numerous!), Yusuf Sobh, Scott Walters, Eric Huber, — Your Name Here —

CHAPTER 2

First Tools for Looking at Data

The single most important question for a working scientist — perhaps the single most useful question anyone can ask — is: "what's going on here?" Answering this question requires creative use of different ways to make pictures of datasets, to summarize them, and to expose whatever structure might be there. This is an activity that is sometimes known as "Descriptive Statistics". There isn't any fixed recipe for understanding a dataset, but there is a rich variety of tools we can use to get insights.

2.1 DATASETS

A dataset is a collection of descriptions of different instances of the same phenomenon. These descriptions could take a variety of forms, but it is important that they are descriptions of the same thing. For example, my grandfather collected the daily rainfall in his garden for many years; we could collect the height of each person in a room; or the number of children in each family on a block; or whether 10 classmates would prefer to be "rich" or "famous". There could be more than one description recorded for each item. For example, when he recorded the contents of the rain gauge each morning, my grandfather could have recorded (say) the temperature and barometric pressure. As another example, one might record the height, weight, blood pressure and body temperature of every patient visiting a doctor's office.

The descriptions in a dataset can take a variety of forms. A description could be **categorical**, meaning that each data item can take a small set of prescribed values. For example, we might record whether each of 100 passers-by preferred to be "Rich" or "Famous". As another example, we could record whether the passers-by are "Male" or "Female". Categorical data could be **ordinal**, meaning that we can tell whether one data item is larger than another. For example, a dataset giving the number of children in a family for some set of families is categorical, because it uses only non-negative integers, but it is also ordinal, because we can tell whether one family is larger than another.

Some ordinal categorical data appears not to be numerical, but can be assigned a number in a reasonably sensible fashion. For example, many readers will recall being asked by a doctor to rate their pain on a scale of 1 to 10 — a question that is usually relatively easy to answer, but is quite strange when you think about it carefully. As another example, we could ask a set of users to rate the usability of an interface in a range from "very bad" to "very good", and then record that using -2 for "very bad", -1 for "bad", 0 for "neutral", 1 for "good", and 2 for "very good".

Many interesting datasets involve **continuous** variables (like, for example, height or weight or body temperature) when you could reasonably expect to encounter any value in a particular range. For example, we might have the heights of all people in a particular room; or the rainfall at a particular place for each day of the year; or the number of children in each family on a list.

You should think of a dataset as a collection of d-tuples (a d-tuple is an ordered list of d elements). Tuples differ from vectors, because we can always add and subtract vectors, but we cannot necessarily add or subtract tuples. We will always write N for the number of tuples in the dataset, and d for the number of elements in each tuple. The number of elements will be the same for every tuple, though sometimes we may not know the value of some elements in some tuples (which means we must figure out how to predict their values, which we will do much later).

Index	net worth
1	100, 360
2	109, 770
3	96, 860
4	97, 860
5	108, 930
6	124, 330
7	101, 300
8	112, 710
9	106, 740
10	120, 170

Index	Taste score	Index	Taste score
1	12.3	11	34.9
2	20.9	12	57.2
3	39	13	0.7
4	47.9	14	25.9
5	5.6	15	54.9
6	25.9	16	40.9
7	37.3	17	15.9
8	21.9	18	6.4
9	18.1	19	18
10	21	20	38.9

TABLE 2.1: *On the* **left**, *net worths of people you meet in a bar, in US \$; I made this data up, using some information from the US Census. The index column, which tells you which data item is being referred to, is usually not displayed in a table because you can usually assume that the first line is the first item, and so on. On the* **right**, *the taste score (I'm not making this up; higher is better) for 20 different cheeses. This data is real (i.e. not made up), and it comes from* *http://lib.stat.cmu.edu/DASL/Datafiles/Cheese.html*.

Each element of a tuple has its own type. Some elements might be categorical. For example, one dataset we shall see several times records entries for Gender; Grade; Age; Race; Urban/Rural; School; Goals; Grades; Sports; Looks; and Money for 478 children, so $d = 11$ and $N = 478$. In this dataset, each entry is categorical data. Clearly, these tuples are not vectors because one cannot add or subtract (say) Genders.

Most of our data will be vectors. We use the same notation for a tuple and for a vector. We write a vector in bold, so $\mathbf{x}$ could represent a vector or a tuple (the context will make it obvious which is intended).

The entire data set is $\{\mathbf{x}\}$. When we need to refer to the i'th data item, we write $\mathbf{x}_i$. Assume we have N data items, and we wish to make a new dataset out of them; we write the dataset made out of these items as $\{\mathbf{x}_i\}$ (the i is to suggest you are taking a set of items and making a dataset out of them).

In this chapter, we will work mainly with continuous data. We will see a variety of methods for plotting and summarizing 1-tuples. We can build these plots from a dataset of d-tuples by extracting the r'th element of each d-tuple.

Mostly, we will deal with continuous data. All through the book, we will see many datasets downloaded from various web sources, because people are so generous about publishing interesting datasets on the web. In the next chapter, we will look at 2-dimensional data, and we look at high dimensional data in chapter **??**.

2.2 WHAT'S HAPPENING? - PLOTTING DATA

The very simplest way to present or visualize a dataset is to produce a table. Tables can be helpful, but aren't much use for large datasets, because it is difficult to get any sense of what the data means from a table. As a continuous example, table 2.1 gives a table of the net worth of a set of people you might meet in a bar (I made this data up). You can scan the table and have a rough sense of what is going on; net worths are quite close to $ 100, 000, and there aren't any very big or very small numbers. This sort of information might be useful, for example, in choosing a bar.

People would like to measure, record, and reason about an extraordinary variety of phenomena. Apparently, one can score the goodness of the flavor of cheese with a number (bigger is better); table 2.1 gives a score for each of thirty cheeses (I did not make up this data, but downloaded it from `http://lib.stat.cmu.edu/DASL/Datafiles/Cheese.html`). You should notice that a few cheeses have very high scores, and most have moderate scores. It's difficult to draw more significant conclusions from the table, though.

Gender	Goal	Gender	Goal
boy	Sports	girl	Sports
boy	Popular	girl	Grades
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	boy	Popular
girl	Popular	girl	Grades
girl	Popular	girl	Sports
girl	Grades	girl	Popular
girl	Sports	girl	Grades
girl	Sports	girl	Sports

TABLE 2.2: *Chase and Dunner (?) collected data on what students thought made other students popular. As part of this effort, they collected information on (a) the gender and (b) the goal of students. This table gives the gender ("boy" or "girl") and the goal (to make good grades —"Grades"; to be popular — "Popular"; or to be good at sports — "Sports"). The table gives this information for the first 20 of 478 students; the rest can be found at `http://lib.stat.cmu.edu/DASL/Datafiles/PopularKids.html`. This data is clearly categorical, and not ordinal.*

Table 2.2 shows a table for a set of categorical data. Psychologists collected data from students in grades 4-6 in three school districts to understand what factors students thought made other students popular. This fascinating data set can be found at `http://lib.stat.cmu.edu/DASL/Datafiles/PopularKids.html`, and was prepared by Chase and Dunner (**?**). Among other things, for each student they asked whether the student's goal was to make good grades ("Grades", for short); to be popular ("Popular"); or to be good at sports ("Sports"). They have this information for 478 students, so a table would be very hard to read. Table 2.2 shows the gender and the goal for the first 20 students in this group. It's rather harder to draw any serious conclusion from this data, because the full table would be so big. We need a more effective tool than eyeballing the table.

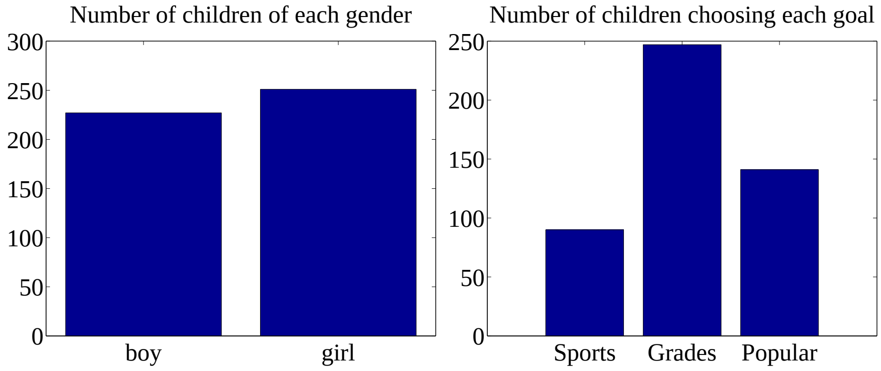

FIGURE 2.1: *On the* **left***, a bar chart of the number of children of each gender in the Chase and Dunner study (). Notice that there are about the same number of boys and girls (the bars are about the same height). On the* **right***, a bar chart of the number of children selecting each of three goals. You can tell, at a glance, that different goals are more or less popular by looking at the height of the bars.*

2.2.1 Bar Charts

A **bar chart** is a set of bars, one per category, where the height of each bar is proportional to the number of items in that category. A glance at a bar chart often exposes important structure in data, for example, which categories are common, and which are rare. Bar charts are particularly useful for categorical data. Figure 2.1 shows such bar charts for the genders and the goals in the student dataset of Chase and Dunner (). You can see at a glance that there are about as many boys as girls, and that there are more students who think grades are important than students who think sports or popularity is important. You couldn't draw either conclusion from Table 2.2, because I showed only the first 20 items; but a 478 item table is very difficult to read.

2.2.2 Histograms

Data is continuous when a data item could take any value in some range or set of ranges. In turn, this means that we can reasonably expect a continuous dataset contains few or no pairs of items that have *exactly* the same value. Drawing a bar chart in the obvious way — one bar per value — produces a mess of unit height bars, and seldom leads to a good plot. Instead, we would like to have fewer bars, each representing more data items. We need a procedure to decide which data items count in which bar.

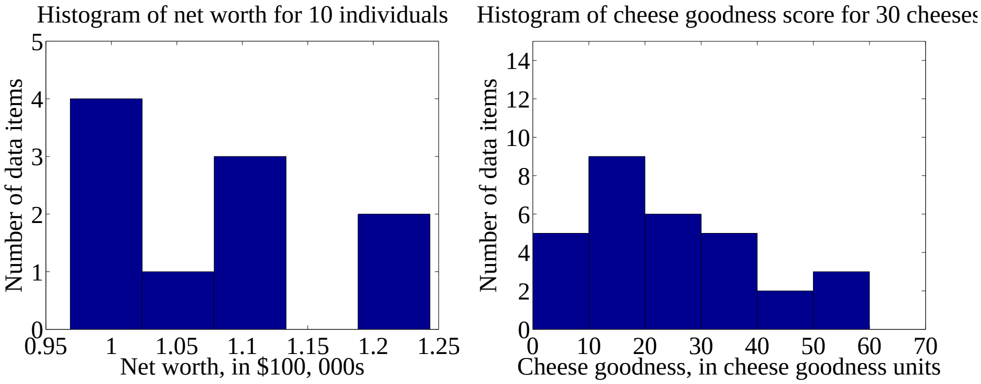

FIGURE 2.2: *On the* **left**, *a histogram of net worths from the dataset described in the text and shown in table 2.1. On the* **right**, *a histogram of cheese goodness scores from the dataset described in the text and shown in table 2.1.*

A simple generalization of a bar chart is a **histogram**. We divide the range of the data into intervals, which do not need to be equal in length. We think of each interval as having an associated pigeonhole, and choose one pigeonhole for each data item. We then build a set of boxes, one per interval. Each box sits on its interval on the horizontal axis, and its height is determined by the number of data items in the corresponding pigeonhole. In the simplest histogram, the intervals that form the bases of the boxes are equally sized. In this case, the height of the box is given by the number of data items in the box.

Figure 2.2 shows a histogram of the data in table 2.1. There are five bars — by my choice; I could have plotted ten bars — and the height of each bar gives the number of data items that fall into its interval. For example, there is one net worth in the range between $102, 500 and $107, 500. Notice that one bar is invisible, because there is no data in that range. This picture suggests conclusions consistent with the ones we had from eyeballing the table — the net worths tend to be quite similar, and around $100, 000.

Figure 2.2 shows a histogram of the data in table 2.1. There are six bars (0-10, 10-20, and so on), and the height of each bar gives the number of data items that fall into its interval — so that, for example, there are 9 cheeses in this dataset whose score is greater than or equal to 10 and less than 20. You can also use the bars to estimate other properties. So, for example, there are 14 cheeses whose score is less than 20, and 3 cheeses with a score of 50 or greater. This picture is much more helpful than the table; you can see at a glance that quite a lot of cheeses have relatively low scores, and few have high scores.

2.2.3 Conditional Histograms

Most people believe that normal body temperature is 98.4^o in Fahrenheit. If you take other people's temperatures often (for example, you might have children), you know that some individuals tend to run a little warmer or a little cooler than this number. I found data giving the body temperature of a set of individuals at

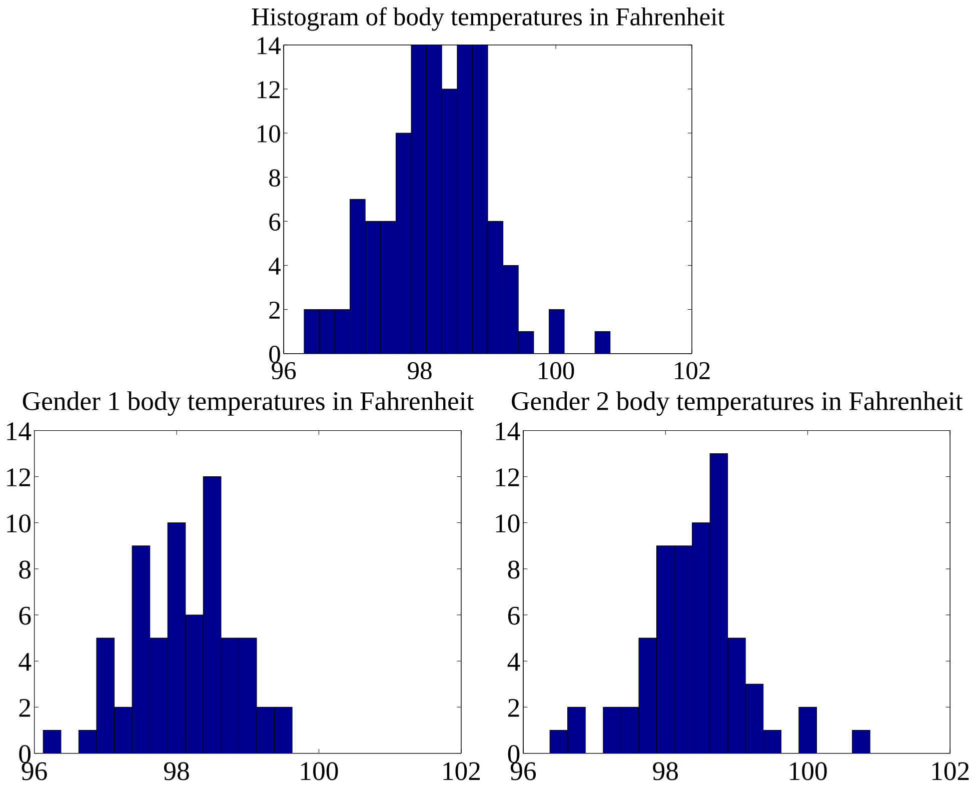

FIGURE 2.3: *On* **top***, a histogram of body temperatures, from the dataset published at* `http://www2.stetson.edu/~jrasp/data.htm`*. These seem to be clustered fairly tightly around one value. The* **bottom row** *shows histograms for each gender (I don't know which is which). It looks as though one gender runs slightly cooler than the other.*

`http://www2.stetson.edu/~jrasp/data.htm`. As you can see from the histogram (figure 2.3), the body temperatures cluster around a small set of numbers. But what causes the variation?

One possibility is gender. We can investigate this possibility by comparing a histogram of temperatures for males with histogram of temperatures for females. Such histograms are sometimes called **conditional histograms** or **class-conditional histograms**, because each histogram is conditioned on something (in this case, the histogram uses only data that comes from gender).

The dataset gives genders (as 1 or 2 - I don't know which is male and which female). Figure 2.3 gives the class conditional histograms. It does seem like individuals of one gender run a little cooler than individuals of the other, although we don't yet have mechanisms to test this possibility in detail (chapter 1).

2.3 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 (section 2.3.1).

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, section 2.3.2).

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 (section 2.3.3).

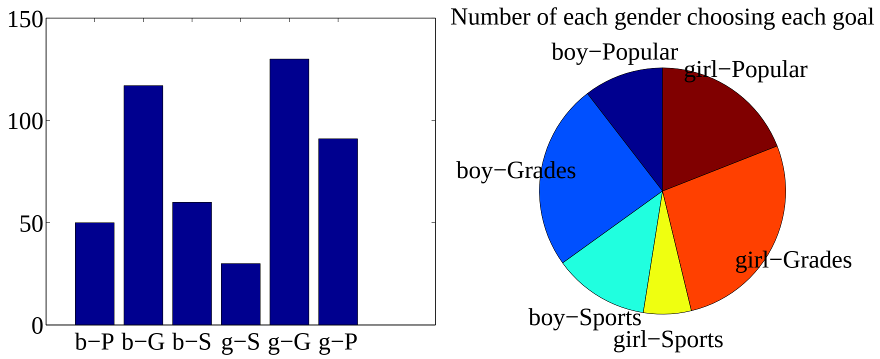

FIGURE 2.4: *I sorted the children in the Chase and Dunner study into six categories (two genders by three goals), and counted the number of children that fell into each cell. I then produced the bar chart on the* **left***, which shows the number of children of each gender, selecting each goal. On the* **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.*

2.3.1 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 2.2, 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 2.4 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 2.4 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 figure 2.4, it's hard to tell that the "boy-sports" category is slightly bigger than the "boy-popular" category (try it; 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 figure 2.4 shows the total number of respondents, and if you refer to figure 2.1, 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.

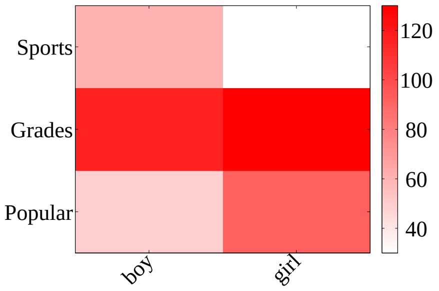

FIGURE 2.5: *A heat map of the Chase and Dunner data. The color of each cell corresponds to the count of the number of elements of that type. The colorbar at the side gives the correspondence between color and count. You can see at a glance that the number of boys and girls who prefer grades is about 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.*

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 figure 2.5

If the categorical data is ordinal, the ordering offers some hints for making

	-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

TABLE 2.3: *I simulated data representing user evaluations of a user interface. 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.*

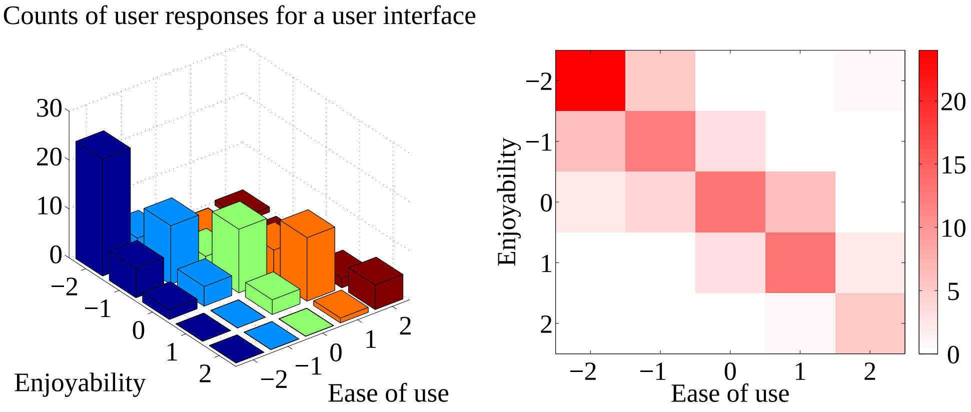

FIGURE 2.6: *On the* **left**, *a 3D bar chart of the data. The height of each bar is given by the number of users in each cell. This figure immediately reveals that users who found the interface hard to use did not like using it either. However, some of the bars at the back are hidden, so some structure 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.*

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 5x5 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 2.3 shows a table and figure 2.6 shows a 3D bar chart for some simulated

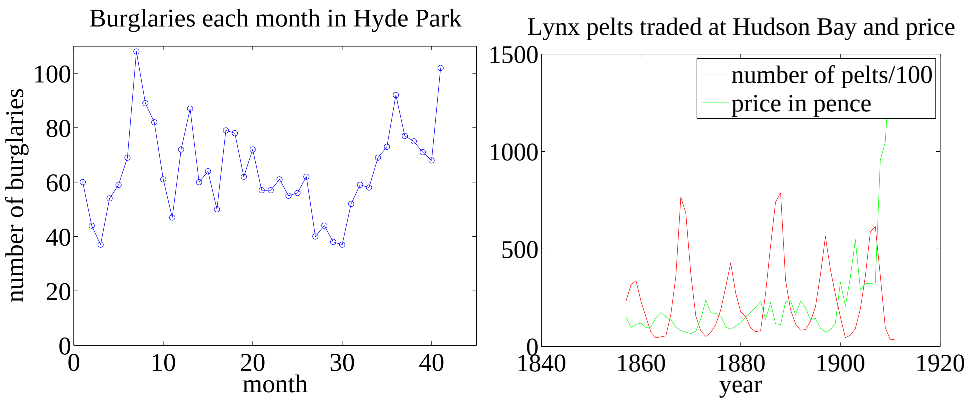

FIGURE 2.7: **Left**, *the number of burglaries in Hyde Park, by month.* **Right**, *a plot of the number of lynx pelts traded at Hudson Bay and of the price paid per pelt, as a function of the year. Notice the scale, and the legend box (the number of pelts is scaled by 100).*

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 (Figure 2.6), another reason they are a good choice.

2.3.2 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 Figure 2.7 (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.

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 figure 2.7. The dataset is famous, because it shows

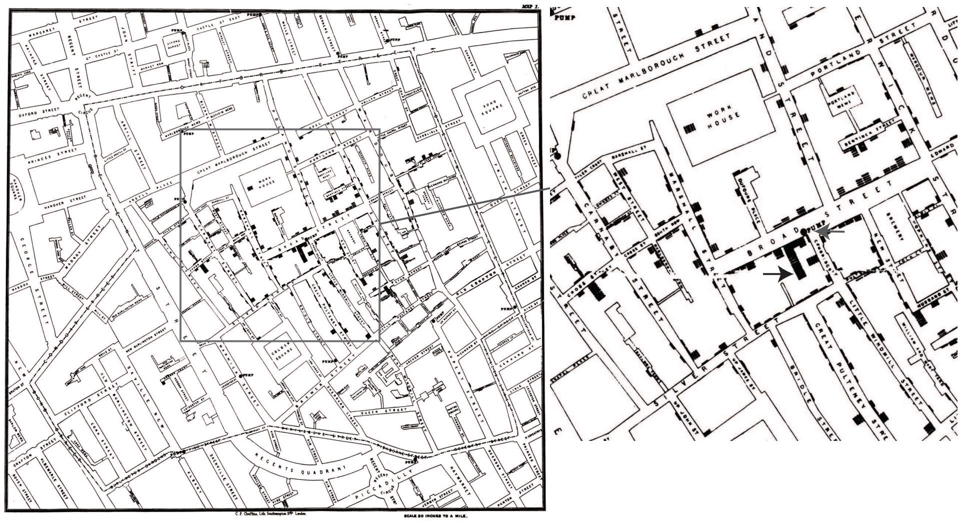

FIGURE 2.8: *Snow's scatter plot of cholera deaths on the* **left***. Each cholera death is plotted as a small bar on the house in which the bar occurred (for example, the black arrow points to one stack of these bars, indicating many deaths, in the detail on the* **right***). Notice the fairly clear 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).*

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.

2.3.3 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

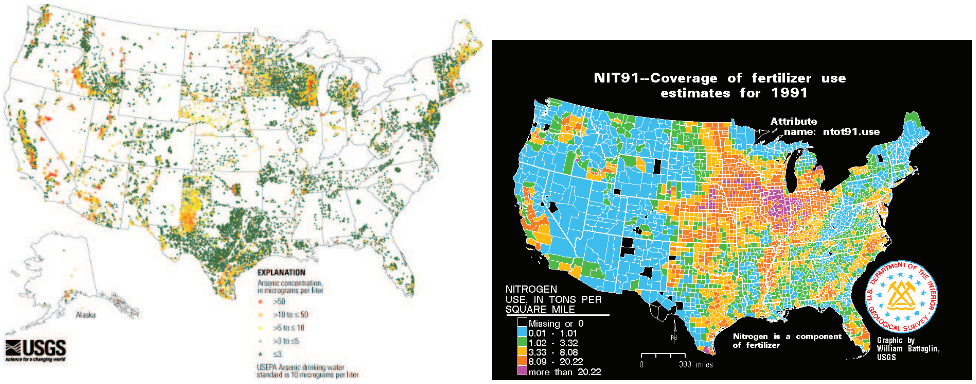
Alaska
USGS
science for a changing world
EXPLANATION
Arsenic concentration, in micrograms per liter
>50
>10 to ≤ 50
>5 to ≤ 10
>3 to ≤5
≤3
USEPA Arsenic drinking water standard is 10 micrograms per liter
NIT91--Coverage of fertilizer use estimates for 1991
Attribute name: ntot91.use
NITROGEN USE, IN TONS PER SQUARE MILE
Missing or 0
0.01 - 1.01
1.02 - 3.32
3.33 - 8.08
8.09 - 20.22
more than 20.22
0
300 miles
Nitrogen is a component of fertilizer
U.S. DEPARTMENT OF THE INTERIOR
GEOLOGICAL SURVEY
Graphic by William Battaglin, USGS

FIGURE 2.9: **Left**, *a scatter plot of arsenic levels in US groundwater, prepared by the US Geological Survey (you can find the data at* `http://water.usgs.gov/GIS/metadata/usgswrd/XML/arsenic_map.xml`*. Here the shape and color of each marker shows the amount of arsenic, and the spatial distribution of the markers shows where the wells were sampled.* **Right**, *the usage of Nitrogen (a component of fertilizer) by US county in 1991, prepared by the US Geological Survey (you can find the data at* `http://water.usgs.gov/GIS/metadata/usgswrd/XML/nit91.xml`*). In this variant of a scatter plot (which usually takes specialized software to prepare) one fills each region with a color indicating the data in that region.*

particular dataset, what we are looking for, how much we are willing to work with complex tools, and our sense of graphic design.

Figure 2.8 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.

Figure 2.9 shows a scatter plot of arsenic levels in groundwater for the United States, prepared by the US Geological Survey. The data set was collected by Focazio and others in 2000; by Welch and others in 2000; and then updated by Ryker 2001. It can be found at `http://water.usgs.gov/GIS/metadata/usgswrd/XML/arsenic_map.xml`. One variant of a scatter plot that is particularly useful for geographic data occurs when one fills regions on a map with different colors, following the data in that region. Figure 2.9 shows the nitrogen usage by US county in 1991; again, this figure was prepared by the US Geological Survey.

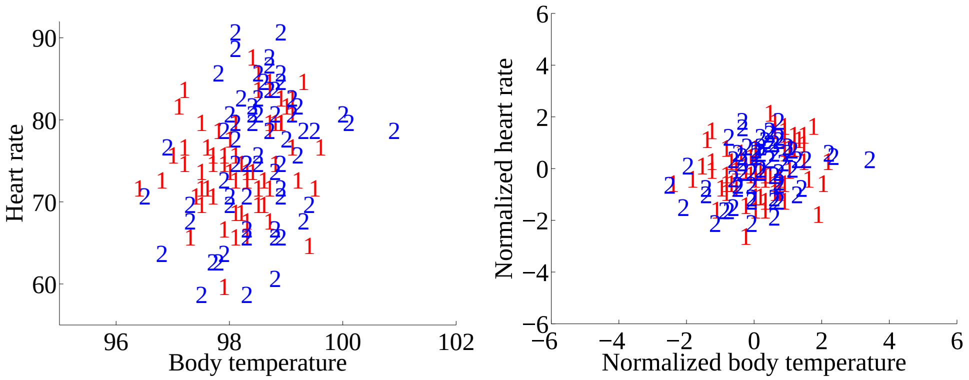

FIGURE 2.10: *A scatter plot of body temperature against heart rate, from the dataset at* `http://www2.stetson.edu/~jrasp/data.htm`*; normtemp.xls. I have separated the two genders by plotting a different symbol for each (though I don't know which gender is indicated by which letter); if you view this 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.*

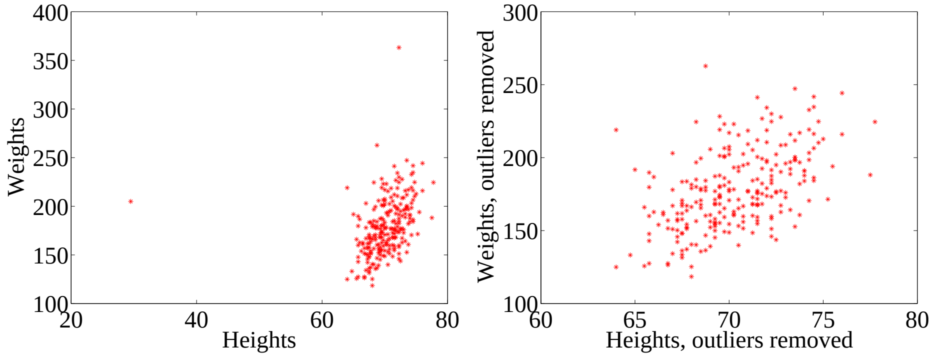

FIGURE 2.11: *A scatter plots of weight against height, from the dataset at* `http://www2.stetson.edu/~jrasp/data.htm`*.* **Left:** *Notice how two outliers dominate the picture, and to show the outliers, the rest of the data has had to be bunched up.* **Right** *shows the data with the outliers removed. The structure is now somewhat clearer.*

2.3.4 Scatter Plots — Scale is a problem

Scatter plots are natural for geographic data, but a scatter plot is a useful, simple tool for ferreting out associations in other kinds of data as well. Now we need some notation. Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. 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,

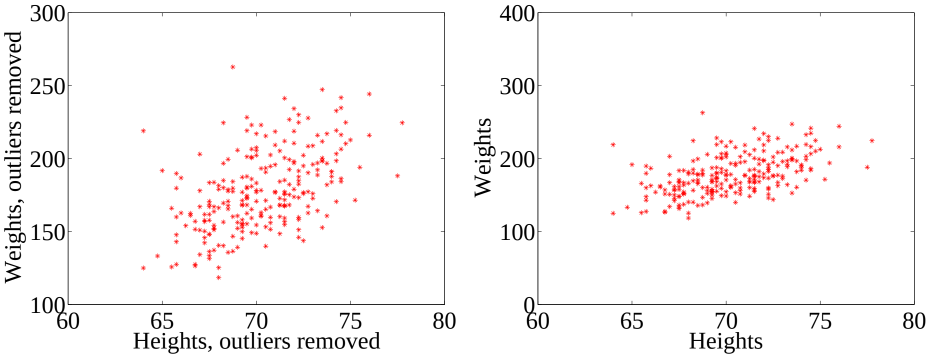

FIGURE 2.12: *Scatter plots of weight against height, from the dataset at* ***http://www2.stetson.edu/~jrasp/data.htm.*** **Left:** *data with two outliers removed, as in figure 2.11.* **Right:** *this data, rescaled slightly. Notice how the data looks less spread out. But there is no difference between the datasets. Instead, your eye is easily confused by a change of scale.*

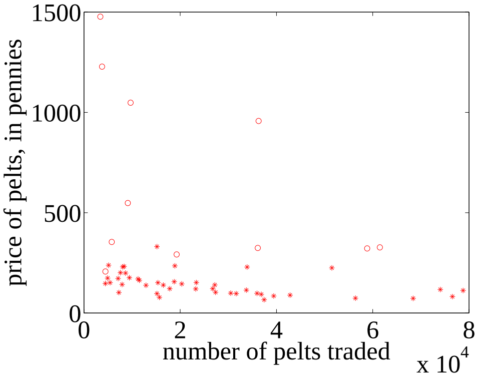

FIGURE 2.13: *A scatter plot of the price of lynx pelts against the number of pelts. I have plotted data for 1901 to the end of the series as circles, and the rest of the data as *'s. It is quite hard to draw any conclusion from this data, because the scale is confusing. Furthermore, the data from 1900 on behaves quite differently from the other data.*

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, figure 2.10 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 2.11 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 2.12 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 figure 2.7, price data before and after 1900 appeared to behave differently. Figure 2.13 shows a scatter plot of the lynx data, where I have plotted number of pelts against price. I 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. The scatter plot does not seem to support the idea that prices go up when supply goes down, which is puzzling, because this is a pretty reliable idea. This turns out to be a scale effect. Scale is an important nuisance, and it's easy to get misled by scale effects.

CHAPTER 3

Summaries and Plots

3.1 SUMMARIZING 1D DATA

For the rest of this chapter, we will assume that data items take values that are continuous real numbers. Furthermore, we will assume that values can be added, subtracted, and multiplied by constants in a meaningful way. Human heights are one example of such data; you can add two heights, and interpret the result as a height (perhaps one person is standing on the head of the other). You can subtract one height from another, and the result is meaningful. You can multiply a height by a constant — say, 1/2 — and interpret the result (A is half as high as B). Not all data is like this. Categorical data is often not like this. For example, you could not add "Grades" to "Popular" in any useful way.

3.1.1 The Mean

One simple and effective summary of a set of data is its **mean**. This is sometimes known as the **average** of the data.

Definition: 3.1 *Mean*

Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. Their mean is

$$\mathsf{mean}\left(\{x\}\right) = \frac{1}{N}\sum_{i=1}^{i=N} x_i.$$

For example, assume you're in a bar, in a group of ten people who like to talk about money. They're average people, and their net worth is given in table 2.1 (you can choose who you want to be in this story). The mean of this data is $107, 903.

An important interpretation of the mean is that it is the best guess of the value of a new data item, given no information at all. In the bar example, if a new person walked into this bar, and you had to guess that person's net worth, you should choose $107, 903.

Properties of the Mean The mean has several important properties you should remember:

- Scaling data scales the mean: or $\mathsf{mean}\left(\{kx_i\}\right) = k\mathsf{mean}\left(\{x_i\}\right)$.
- Translating data translates the mean: or $\mathsf{mean}\left(\{x_i + c\}\right) = \mathsf{mean}\left(\{x_i\}\right) + c$.

- The sum of signed differences from the mean is zero. This means that

$$\sum_{i=1}^{N}(x_i - \mathsf{mean}(\{x_i\})) = 0.$$

- Choose the number μ such that the sum of squared distances of data points to μ is minimized. That number is the mean. In notation

$$\underset{\mu}{\text{arg min}} \sum_i (x_i - \mu)^2 = \mathsf{mean}(\{x_i\})$$

These properties are easy to prove (and so easy to remember). All but one proof is relegated to the exercises.

Proposition: $\underset{\mu}{\mathit{arg\ min}} \sum_i (x_i - \mu)^2 = \mathsf{mean}(\{x\})$

Proof: Choose the number μ such that the sum of squared distances of data points to μ is minimized. That number is the mean. In notation:

$$\underset{\mu}{\text{arg min}} \sum_i (x_i - \mu)^2 = \mathsf{mean}(\{x\})$$

We can show this by actually minimizing the expression. We must have that the derivative of the expression we are minimizing is zero at the value of μ we are seeking. So we have

$$\begin{aligned}\frac{d}{d\mu}\sum_{i=1}^{N}(x_i - \mu)^2 &= \sum_{i=1}^{N} 2(x_i - \mu) \\ &= 2\sum_{i=1}^{N}(x_i - \mu) \\ &= 0\end{aligned}$$

so that $2N\mathsf{mean}(\{x\}) - 2N\mu = 0$, which means that $\mu = \mathsf{mean}(\{x\})$.

Property 3.1: The Average Squared Distance to the Mean is Minimized

3.1.2 Standard Deviation and Variance

We would also like to know the extent to which data items are close to the mean. This information is given by the **standard deviation**, which is the root mean square of the offsets of data from the mean.

Definition: 3.2 *Standard deviation*

Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. The standard deviation of this dataset is is:

$$\mathsf{std}\left(x_i\right) = \sqrt{\frac{1}{N}\sum_{i-1}^{i=N}(x_i - \mathsf{mean}\left(\{x\}\right))^2} = \sqrt{\mathsf{mean}\left(\{(x_i - \mathsf{mean}\left(\{x\}\right))^2\}\right)}.$$

You should think of the standard deviation as a scale. It measures the size of the average deviation from the mean for a dataset. When the standard deviation of a dataset is large, there are many items with values much larger than, or much smaller than, the mean. When the standard deviation is small, most data items have values close to the mean. This means it is helpful to talk about how many standard devations away from the mean a particular data item is. Saying that data item x_j is "within k standard deviations from the mean" means that

$$\mathsf{abs}\left(x_j - \mathsf{mean}\left(\{x\}\right)\right) \leq k\mathsf{std}\left(x_i\right).$$

Similarly, saying that data item x_j is "more than k standard deviations from the mean" means that

$$\mathsf{abs}\left(x_i - \mathsf{mean}\left(\{x\}\right)\right) > k\mathsf{std}\left(x\right).$$

As I will show below, there must be some data at least one standard deviation away from the mean, and there can be very few data items that are many standard deviations away from the mean.

Properties of the Standard Deviation Standard deviation has very important properties:

- Translating data does not change the standard deviation, i.e. $\mathsf{std}\left(x_i + c\right) = \mathsf{std}\left(x_i\right)$.
- Scaling data scales the standard deviation, i.e. $\mathsf{std}\left(kx_i\right) = k\mathsf{std}\left(x_i\right)$.
- For any dataset, there can be only a few items that are many standard deviations away from the mean. In particular, assume we have N data items, x_i, whose standard deviation is σ. Then there are at most $\frac{1}{k^2}$ data points lying k or more standard deviations away from the mean.
- For any dataset, there must be at least one data item that is at least one standard deviation away from the mean.

The first two properties are easy to prove, and are relegated to the exercises.

Proposition: *Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. Assume the standard deviation of this dataset is $\mathsf{std}(x) = \sigma$. Then there are at most $\frac{1}{k^2}$ data points lying k or more standard deviations away from the mean.*

Proof: Assume the mean is zero. There is no loss of generality here, because translating data translates the mean, but doesn't change the standard deviation. The way to prove this is to construct a dataset with the largest possible fraction r of data points lying k or more standard deviations from the mean. To achieve this, our data should have $N(1-r)$ data points each with the value 0, because these contribute 0 to the standard deviation. It should have Nr data points with the value $k\sigma$; if they are further from zero than this, each will contribute more to the standard deviation, so the fraction of such points will be fewer. Because

$$\mathsf{std}(x) = \sigma = \sqrt{\frac{\sum_i x_i^2}{N}}$$

we have that, for this rather specially constructed dataset,

$$\sigma = \sqrt{\frac{Nrk^2\sigma^2}{N}}$$

so that

$$r = \frac{1}{k^2}.$$

We constructed the dataset so that r would be as large as possible, so

$$r \geq \frac{1}{k^2}$$

for any kind of data at all.

Property 3.2: For any dataset, it is hard for data items to get many standard deviations away from the mean.

The bound of box 3.1.2 is true for *any kind of data.* This bound implies that, for example, at most 100% of *any* dataset could be one standard deviation away from the mean, 25% of *any* dataset is 2 standard deviations away from the mean and at most 11% of *any* dataset could be 3 standard deviations away from the mean. But the configuration of data that achieves this bound is very unusual. This means the bound tends to wildly overstate how much data is far from the mean for most practical datasets. Most data has more random structure, meaning that we expect to see very much *less* data far from the mean than the bound predicts. For example, much data can reasonably be modelled as coming from a normal distribution (a topic we'll go into later). For such data, we expect that about 68% of the data is within one standard deviation of the mean, 95% is within two standard deviations of the mean, and 99.7% is within three standard deviations of the mean, and the percentage of data that is within ten standard deviations of the mean is essentially indistinguishable from 100%. This kind of behavior is quite common; the crucial point about the standard deviation is that you won't see much data that lies many standard deviations from the mean, because you can't.

Proposition: $(\mathsf{std}\,(x))^2 \leq \max_i (x_i - \mathsf{mean}\,(\{x\}))^2$.

Proof: You can see this by looking at the expression for standard deviation. We have

$$\mathsf{std}\,(x) = \sqrt{\frac{1}{N}\sum_{i-1}^{i=N}(x_i - \mathsf{mean}\,(\{x\}))^2}.$$

Now, this means that

$$N(\mathsf{std}\,(x))^2 = \sum_{i-1}^{i=N}(x_i - \mathsf{mean}\,(\{x\}))^2.$$

But

$$\sum_{i-1}^{i=N}(x_i - \mathsf{mean}\,(\{x\}))^2 \leq N \max_i (x_i - \mathsf{mean}\,(\{x\}))^2$$

so

$$(\mathsf{std}\,(x))^2 \leq \max_i (x_i - \mathsf{mean}\,(\{x\}))^2.$$

Property 3.3: For any dataset, there must be at least one data item that is at least one standard deviation away from the mean.

Boxes 3.1.2 and 3.1.2 mean that the standard deviation is quite informative. Very little data is many standard deviations away from the mean; similarly, at least some of the data should be one or more standard deviations away from the mean. So the standard deviation tells us how data points are scattered about the mean.

Potential point of confusion: There is an ambiguity that comes up often here because two (very slightly) different numbers are called the standard deviation of a dataset. One — the one we use in this chapter — is an estimate of the scale of the data, as we describe it. The other differs from our expression very slightly; one computes

$$\sqrt{\frac{\sum_i (x_i - \mathsf{mean}\,(\{x\}))^2}{N-1}}$$

(notice the $N-1$ for our N). If N is large, this number is basically the same as the number we compute, but for smaller N there is a difference that can be significant. Irritatingly, this number is also called the standard deviation; even more irritatingly, we will have to deal with it, but not yet. I mention it now because you may look up terms I have used, find this definition, and wonder. Don't worry - the N in our expressions is the right thing to use for what we're doing.

3.1.3 Variance

It turns out that thinking in terms of the square of the standard deviation, which is known as the **variance**, will allow us to generalize our summaries to apply to higher dimensional data.

Definition: 3.3 *Variance*

Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. where $N > 1$. Their variance is:

$$\mathsf{var}\left(\{x\}\right) = \frac{1}{N}\left(\sum_{i-1}^{i=N}(x_i - \mathsf{mean}\left(\{x\}\right))^2\right) = \mathsf{mean}\left(\left\{(x_i - \mathsf{mean}\left(\{x\}\right))^2\right\}\right).$$

One good way to think of the variance is as the mean-square error you would incur if you replaced each data item with the mean. Another is that it is the square of the standard deviation.

Properties of the Variance The properties of the variance follow from the fact that it is the square of the standard deviation. We have that:

- Translating data does not change the variance, i.e. $\mathsf{var}\left(\{x + c\}\right) = \mathsf{var}\left(\{x\}\right)$.
- Scaling data scales the variance by a square of the scale, i.e. $\mathsf{var}\left(\{kx\}\right) = k^2\mathsf{var}\left(\{x\}\right)$.

While one could restate the other two properties of the standard deviation in terms of the variance, it isn't really natural to do so. The standard deviation is in the same units as the original data, and should be thought of as a scale. Because the variance is the square of the standard deviation, it isn't a natural scale (unless you take its square root!).

3.1.4 The Median

One problem with the mean is that it can be affected strongly by extreme values. Go back to the bar example, of section 3.1.1. Now Warren Buffett (or Bill Gates, or your favorite billionaire) walks in. What happened to the average net worth?

Assume your billionaire has net worth \$ 1, 000, 000, 000. Then the mean net worth suddenly has become

$$\frac{10 \times \$107,903 + \$1,000,000,000}{11} = \$91,007,184$$

But this mean isn't a very helpful summary of the people in the bar. It is probably more useful to think of the net worth data as ten people together with one billionaire. The billionaire is known as an **outlier**.

One way to get outliers is that a small number of data items are very different, due to minor effects you don't want to model. Another is that the data was misrecorded, or mistranscribed. Another possibility is that there is just too much variation in the data to summarize it well. For example, a small number of extremely wealthy people could change the average net worth of US residents dramatically, as the example shows. An alternative to using a mean is to use a **median**.

> **Definition: 3.4** *Median*
>
> The median of a set of data points is obtained by sorting the data points, and finding the point halfway along the list. If the list is of even length, it's usual to average the two numbers on either side of the middle. We write
>
> $$\mathsf{median}\left(\{x_i\}\right)$$
>
> for the operator that returns the median.

For example,

$$\mathsf{median}\left(\{3, 5, 7\}\right) = 5,$$

$$\mathsf{median}\left(\{3, 4, 5, 6, 7\}\right) = 5,$$

and

$$\mathsf{median}\left(\{3, 4, 5, 6\}\right) = 4.5.$$

For much, but not all, data, you can expect that roughly half the data is smaller than the median, and roughly half is larger than the median. Sometimes this property fails. For example,

$$\mathsf{median}\left(\{1, 2, 2, 2, 2, 2, 2, 2, 3\}\right) = 2.$$

With this definition, the median of our list of net worths is \$107, 835. If we insert the billionaire, the median becomes \$108, 930. Notice by how little the number has changed — it remains an effective summary of the data.

Properties of the median You can think of the median of a dataset as giving the "middle" or "center" value. This means it is rather like the mean, which also gives a (slightly differently defined) "middle" or "center" value. The mean has the important properties that if you translate the dataset, the mean translates, and if you scale the dataset, the mean scales. The median has these properties, too:

- Translating data translates the median, i.e. $\mathsf{median}\left(\{x + c\}\right) = \mathsf{median}\left(\{x\}\right) + c$.
- Scaling data scales the median by the same scale, i.e. $\mathsf{median}\left(\{kx\}\right) = k\mathsf{median}\left(\{x\}\right)$.

Each is easily proved, and proofs are relegated to the exercises.

3.1.5 Interquartile Range

Outliers can affect standard deviations severely, too. For our net worth data, the standard deviation without the billionaire is \$9265, but if we put the billionaire in there, it is $\$3.014 \times 10^8$. When the billionaire is in the dataset, all but one of the data items lie about a third of a standard deviation away from the mean; the other one (the billionaire) is many standard deviations away from the mean. In this case, the standard deviation has done its work of informing us that there are huge changes in the data, but isn't really helpful.

The problem is this: describing the net worth data with billionaire as a having a mean of $\$9.101 \times 10^7$ with a standard deviation of $\$3.014 \times 10^8$ really isn't terribly helpful. Instead, the data really should be seen as a clump of values that are near $\$100,000$ and moderately close to one another, and one massive number (the billionaire outlier).

One thing we could do is simply remove the billionaire and compute mean and standard deviation. This isn't always easy to do, because it's often less obvious which points are outliers. An alternative is to follow the strategy we did when we used the median. Find a summary that describes scale, but is less affected by outliers than the standard deviation. This is the **interquartile range**; to define it, we need to define percentiles and quartiles, which are useful anyway.

Definition: 3.5 *Percentile*

The k'th percentile is the value such that $k\%$ of the data is less than or equal to that value. We write $\mathsf{percentile}(\{x\}, k)$ for the k'th percentile of dataset $\{x\}$.

Definition: 3.6 *Quartiles*

The first quartile of the data is the value such that 25% of the data is less than or equal to that value (i.e. $\mathsf{percentile}(\{x\}, 25)$). The second quartile of the data is the value such that 50% of the data is less than or equal to that value, which is usually the median (i.e. $\mathsf{percentile}(\{x\}, 50)$). The third quartile of the data is the value such that 75% of the data is less than or equal to that value (i.e. $\mathsf{percentile}(\{x\}, 75)$).

Definition: 3.7 *Interquartile Range*

The interquartile range of a dataset $\{x\}$ is $\mathsf{iqr}\{x\} = \mathsf{percentile}(\{x\}, 75) - \mathsf{percentile}(\{x\}, 25)$.

Like the standard deviation, the interquartile range gives an estimate of how widely the data is spread out. But it is quite well-behaved in the presence of outliers. For our net worth data without the billionaire, the interquartile range is $12350; with the billionaire, it is $17710.

Properties of the interquartile range You can think of the interquartile range of a dataset as giving an estimate of the scale of the difference from the mean. This means it is rather like the standard deviation, which also gives a (slightly differently defined) scale. The standard deviation has the important properties that if you translate the dataset, the standard deviation translates, and if you scale the dataset, the standard deviation scales. The interquartile range has these properties, too:

- Translating data does not change the interquartile range, i.e. $\mathsf{iqr}\{x+c\} = \mathsf{iqr}\{x\}$.
- Scaling data scales the interquartile range by the same scale, i.e. $\mathsf{iqr}\{kx\} = k^2\mathsf{iqr}\{x\}$.

Each is easily proved, and proofs are relegated to the exercises.

3.1.6 Using Summaries Sensibly

One should be careful how one summarizes data. For example, the statement that "the average US family has 2.6 children" invites mockery (the example is from Andrew Vickers' book *What is a p-value anyway?*), because you can't have fractions of a child — no family has 2.6 children. A more accurate way to say things might be "the average of the number of children in a US family is 2.6", but this is clumsy. What is going wrong here is the 2.6 is a mean, but the number of children in a family is a categorical variable. Reporting the mean of a categorical variable is often a bad idea, because you may never encounter this value (the 2.6 children). For a categorical variable, giving the median value and perhaps the interquartile range often makes much more sense than reporting the mean.

For continuous variables, reporting the mean is reasonable because you could expect to encounter a data item with this value, even if you haven't seen one in the particular data set you have. It is sensible to look at both mean and median; if they're significantly different, then there is probably something going on that is worth understanding. You'd want to plot the data using the methods of the next section before you decided what to report.

You should also be careful about how precisely numbers are reported (equivalently, the number of significant figures). Numerical and statistical software will produce very large numbers of digits freely, but not all are always useful. This is a particular nuisance in the case of the mean, because you might add many numbers, then divide by a large number; in this case, you will get many digits, but some might not be meaningful. For example, Vickers (ibid) describes a paper reporting the mean length of pregnancy as 32.833 weeks. That fifth digit suggests we know the mean length of pregnancy to about 0.001 weeks, or roughly 10 minutes. Neither medical interviewing nor people's memory for past events is that detailed. Furthermore, when you interview them about embarrassing topics, people quite often lie. There is no prospect of knowing this number with this precision.

People regularly report silly numbers of digits because it is easy to miss the harm caused by doing so. But the harm is there: you are implying to other people, and to yourself, that you know something more accurately than you do. At some point, someone will suffer for it.

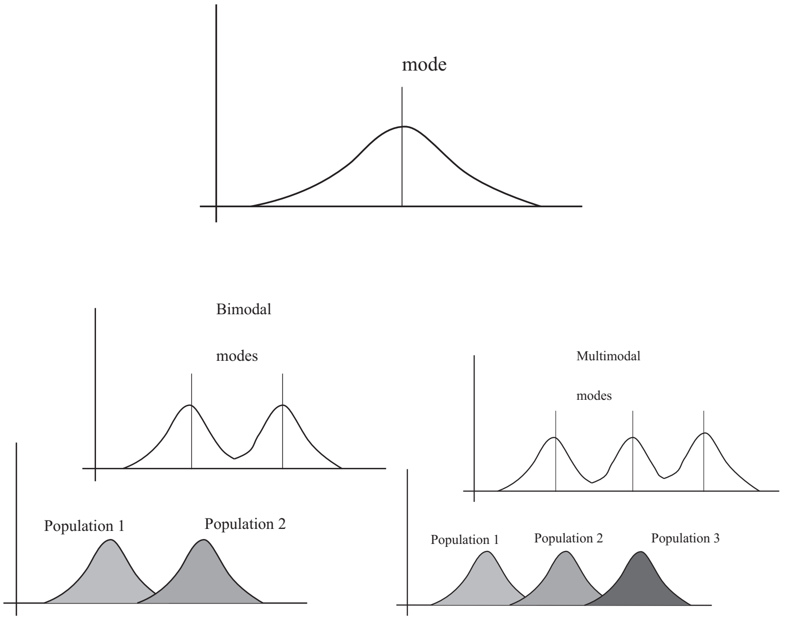

FIGURE 3.1: *Many histograms are unimodal, like the example on the* **top***; there is one peak, or mode. Some are bimodal (two peaks;* **bottom left***) or even multimodal (two or more peaks;* **bottom right***). One common reason (but not the only reason) is that there are actually two populations being conflated in the histograms. For example, measuring adult heights might result in a bimodal histogram, if male and female heights were slightly different. As another example, measuring the weight of dogs might result in a multimodal histogram if you did not distinguish between breeds (eg chihauhau, terrier, german shepherd, pyranean mountain dog, etc.).*

3.2 PLOTS AND SUMMARIES

Knowing the mean, standard deviation, median and interquartile range of a dataset gives us some information about what its histogram might look like. In fact, the summaries give us a language in which to describe a variety of characteristic properties of histograms that are worth knowing about (Section 3.2.1). Quite remarkably, many different datasets have the same shape of histogram (Section 3.2.2). For such data, we know roughly what percentage of data items are how far from the mean.

Complex datasets can be difficult to interpret with histograms alone, because it is hard to compare many histograms by eye. Section 3.2.3 describes a clever plot of various summaries of datasets that makes it easier to compare many cases.

3.2.1 Some Properties of Histograms

The **tails** of a histogram are the relatively uncommon values that are significantly larger (resp. smaller) than the value at the peak (which is sometimes called the **mode**). A histogram is **unimodal** if there is only one peak; if there are more than one, it is **multimodal**, with the special term **bimodal** sometimes being used for

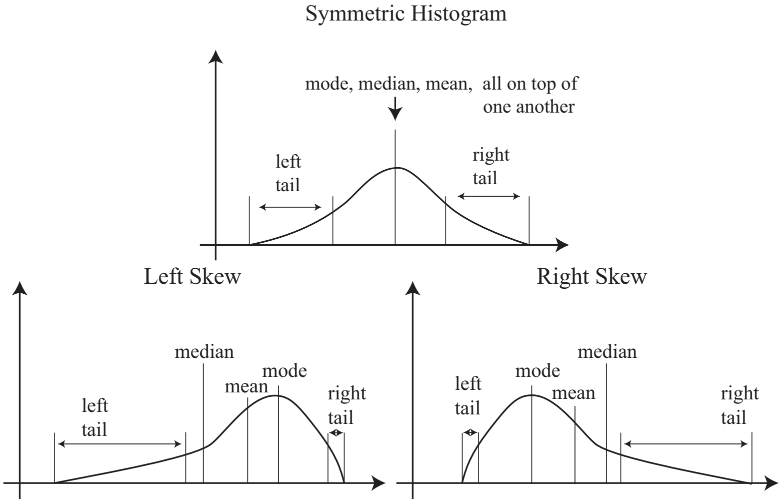

FIGURE 3.2: *On the* **top**, *an example of a symmetric histogram, showing its tails (relatively uncommon values that are significantly larger or smaller than the peak or mode).* **Lower left**, *a sketch of a left-skewed histogram. Here there are few large values, but some very small values that occur with significant frequency. We say the left tail is "long", and that the histogram is left skewed (confusingly, this means the main bump is to the right).* **Lower right**, *a sketch of a right-skewed histogram. Here there are few small values, but some very large values that occur with significant frequency. We say the right tail is "long", and that the histogram is right skewed (confusingly, this means the main bump is to the left).*

the case where there are two peaks (Figure 3.1). The histograms we have seen have been relatively symmetric, where the left and right tails are about as long as one another. Another way to think about this is that values a lot larger than the mean are about as common as values a lot smaller than the mean. Not all data is symmetric. In some datasets, one or another tail is longer (figure 3.2). This effect is called **skew**.

Skew appears often in real data. SOCR (the Statistics Online Computational Resource) publishes a number of datasets. Here we discuss a dataset of citations to faculty publications. For each of five UCLA faculty members, SOCR collected the number of times each of the papers they had authored had been cited by other authors (data at `http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_072108_H_Index_Pubs`). Generally, a small number of papers get many citations, and many papers get few citations. We see this pattern in the histograms of citation numbers (figure 3.3). These are very different from (say) the body temperature pictures. In the citation histograms, there are many data items that have very few citations, and few that have many citations. This means that the right tail of the histogram is longer, so the histogram is skewed to the right.

One way to check for skewness is to look at the histogram; another is to compare mean and median (though this is not foolproof). For the first citation

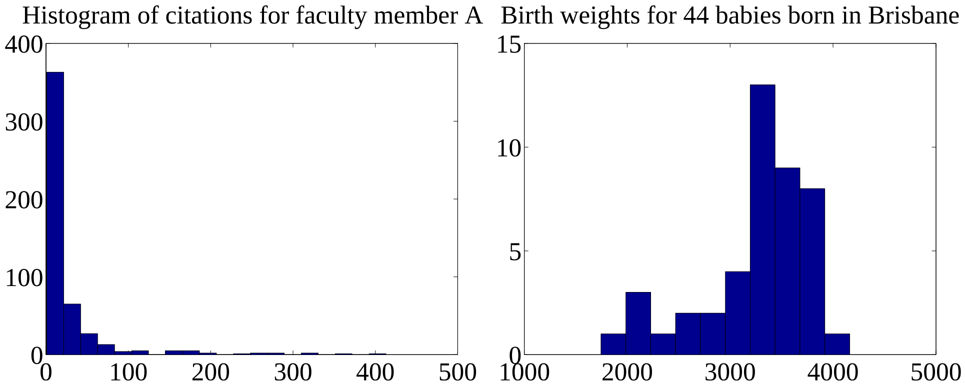

FIGURE 3.3: *On the* **left**, *a histogram of citations for a faculty member, from data at* ***http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_072108_H_Index_Pubs****. Very few publications have many citations, and many publications have few. This means the histogram is strongly right-skewed. On the* **right**, *a histogram of birth weights for 44 babies borne in Brisbane in 1997. This histogram looks slightly left-skewed.*

histogram, the mean is 24.7 and the median is 7.5; for the second, the mean is 24.4, and the median is 11. In each case, the mean is a lot bigger than the median. Recall the definition of the median (form a ranked list of the data points, and find the point halfway along the list). For much data, the result is larger than about half of the data set and smaller than about half the dataset. So if the median is quite small compared to the mean, then there are many small data items and a small number of data items that are large — the right tail is longer, so the histogram is skewed to the right.

Left-skewed data also occurs; figure 3.3 shows a histogram of the birth weights of 44 babies born in Brisbane, in 1997 (from `http://www.amstat.org/publications/jse/jse_data_archive.htm`). This data appears to be somewhat left-skewed, as birth weights can be a lot smaller than the mean, but tend not to be much larger than the mean.

Skewed data is often, but not always, the result of constraints. For example, good obstetrical practice tries to ensure that very large birth weights are rare (birth is typically induced before the baby gets too heavy), but it may be quite hard to avoid some small birth weights. This could could skew birth weights to the left (because large babies will get born, but will not be as heavy as they could be if obstetricians had not interfered). Similarly, income data can be skewed to the right by the fact that income is always positive. Test mark data is often skewed — whether to right or left depends on the circumstances — by the fact that there is a largest possible mark and a smallest possible mark.

3.2.2 Standard Coordinates and Normal Data

It is useful to look at lots of histograms, because it is often possible to get some useful insights about data. However, in their current form, histograms are hard to compare. This is because each is in a different set of units. A histogram for length data will consist of boxes whose horizontal units are, say, metres; a histogram for mass data will consist of boxes whose horizontal units are in, say, kilograms. Furthermore, these histograms typically span different ranges.

We can make histograms comparable by (a) estimating the "location" of the plot on the horizontal axis and (b) estimating the "scale" of the plot. The location is given by the mean, and the scale by the standard deviation. We could then normalize the data by subtracting the location (mean) and dividing by the standard deviation (scale). The resulting values are unitless, and have zero mean. They are often known as **standard coordinates**.

Definition: 3.8 *Standard coordinates*

Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. We represent these data items in standard coordinates by computing

$$\hat{x}_i = \frac{(x_i - \mathsf{mean}\left(\{x\}\right))}{\mathsf{std}\left(x\right)}.$$

We write $\{\hat{x}\}$ for a dataset that happens to be in standard coordinates.

Standard coordinates have some important properties. Assume we have N data items. Write x_i for the i'th data item, and $\hat{x}_i$ for the i'th data item in standard coordinates (I sometimes refer to these as "normalized data items"). Then we have

$$\mathsf{mean}\left(\{\hat{x}\}\right) = 0.$$

We also have that

$$\mathsf{std}\left(\hat{x}\right) = 1.$$

An extremely important fact about data is that, for many kinds of data, histograms of these standard coordinates look the same. Many completely different datasets produce a histogram that, in standard coordinates, has a very specific appearance. It is symmetric, unimodal; it looks like a narrow bump. If there were enough data points and the histogram boxes were small enough, the curve would look like the curve in figure 3.4. This phenomenon is so important that data of this form has a special name.

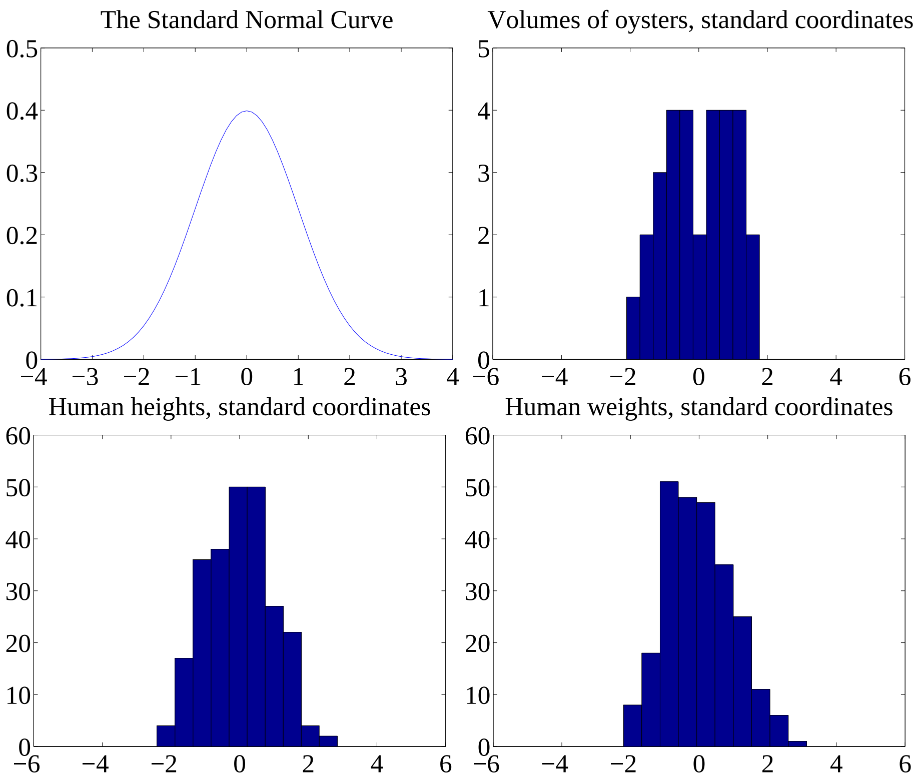

FIGURE 3.4: *Data is standard normal data when its histogram takes a stylized, bell-shaped form, plotted above. One usually requires a lot of data and very small histogram boxes for this form to be reproduced closely. Nonetheless, the histogram for normal data is unimodal (has a single bump) and is symmetric; the tails fall off fairly fast, and there are few data items that are many standard deviations from the mean. Many quite different data sets have histograms that are similar to the normal curve; I show three such datasets here.*

Definition: 3.9 *Standard normal data*

Data is **standard normal data** if, when we have a great deal of data, the histogram is a close approximation to the **standard normal curve**. This curve is given by

$$y(x) = \frac{1}{\sqrt{2pi}} e^{\left(-x^2/2\right)}$$

(which is shown in figure 3.4).

Definition: 3.10 *Normal data*

Data is **normal data** if, when we subtract the mean and divide by the standard deviation (i.e. compute standard coordinates), it becomes standard normal data.

It is not always easy to tell whether data is normal or not, and there are a variety of tests one can use, which we discuss later. However, there are many examples of normal data. Figure 3.4 shows a diverse variety of data sets, plotted as histograms in standard coordinates. These include: the volumes of 30 oysters (from `http://www.amstat.org/publications/jse/jse_data_archive.htm`; look for 30oysters.dat.txt); human heights (from `http://www2.stetson.edu/~jrasp/data.htm`; look for bodyfat.xls, with two outliers removed); and human weights (from `http://www2.stetson.edu/~jrasp/data.htm`; look for bodyfat.xls, with two outliers removed).

Properties of normal data For the moment, assume we know that a dataset is normal. Then we expect it to have the following properties:

- If we normalize it, its histogram will be close to the standard normal curve. This means, among other things, that the data is not significantly skewed.
- About 68% of the data lie within one standard deviation of the mean. We will prove this later.
- About 95% of the data lie within two standard deviations of the mean. We will prove this later.
- About 99% of the data lie within three standard deviations of the mean. We will prove this later.

In turn, these properties imply that data that contains outliers (points many standard deviations away from the mean) is not normal. This is usually a very safe assumption. It is quite common to model a dataset by excluding a small number of outliers, then modelling the remaining data as normal. For example, if I exclude two outliers from the height and weight data from `http://www2.stetson.edu/~jrasp/data.htm`, the data looks pretty close to normal.

3.2.3 Boxplots

It is usually hard to compare multiple histograms by eye. One problem with comparing histograms is the amount of space they take up on a plot, because each histogram involves multiple vertical bars. This means it is hard to plot multiple overlapping histograms cleanly. If you plot each one on a separate figure, you have to handle a large number of separate figures; either you print them too small to see enough detail, or you have to keep flipping over pages.

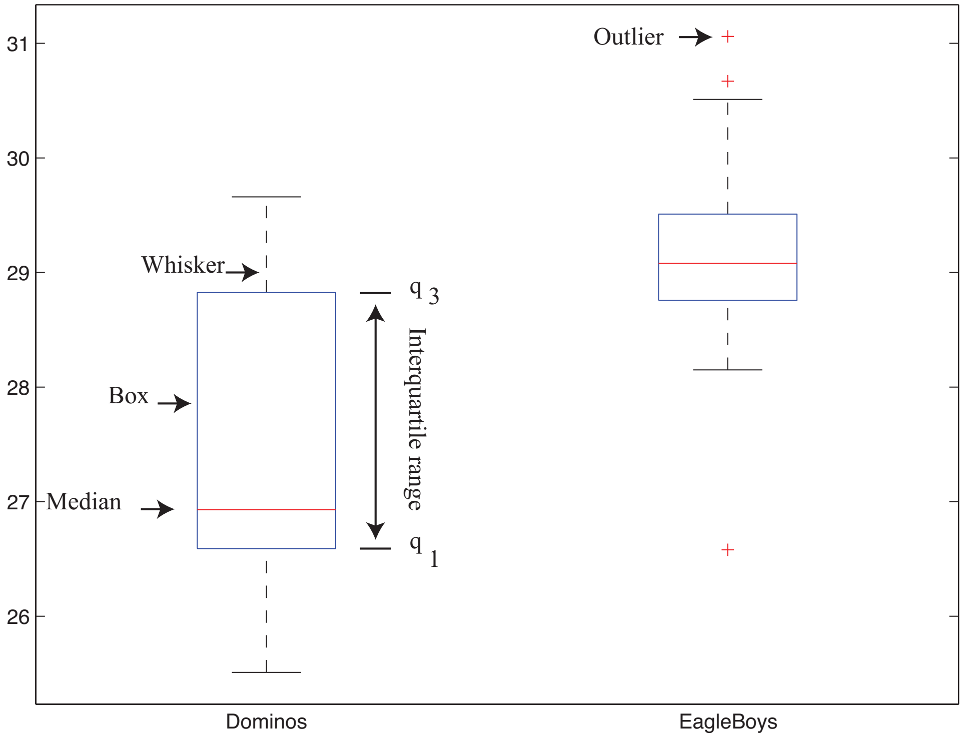

FIGURE 3.5: *A boxplot showing the box, the median, the whiskers and two outliers. Notice that we can compare the two datasets rather easily; the next section explains the comparison.*

A **boxplot** is a way to plot data that simplifies comparison. A boxplot displays a dataset as a vertical picture. There is a vertical box whose height corresponds to the interquartile range of the data (the width is just to make the figure easy to interpret). Then there is a horizontal line for the median; and the behavior of the rest of the data is indicated with whiskers and/or outlier markers. This means that each dataset makes is represented by a vertical structure, making it easy to show multiple datasets on one plot *and interpret the plot* (Figure 3.5).

To build a boxplot, we first plot a box that runs from the first to the third quartile. We then show the median with a horizontal line. We then decide which data items should be outliers. A variety of rules are possible; for the plots I show, I used the rule that data items that are larger than $q_3 + 1.5(q_3 - q_1)$ or smaller than $q_1 - 1.5(q_3 - q_1)$, are outliers. This criterion looks for data items that are more than one and a half interquartile ranges above the third quartile, or more than one and a half interquartile ranges below the first quartile.

Once we have identified outliers, we plot these with a special symbol (crosses in the plots I show). We then plot whiskers, which show the range of non-outlier data. We draw a whisker from q_1 to the smallest data item that is not an outlier, and from q_3 to the largest data item that is not an outlier. While all this sounds

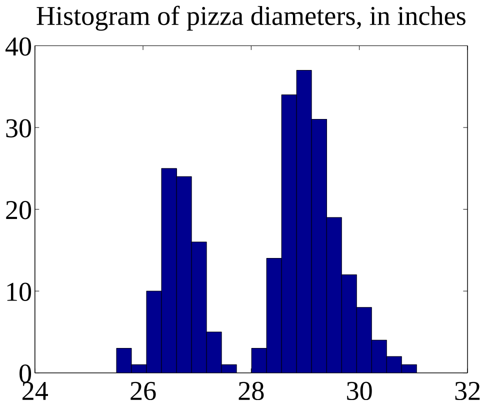

FIGURE 3.6: *A histogram of pizza diameters from the dataset described in the text. Notice that there seem to be two populations.*

complicated, any reasonable programming environment will have a function that will do it for you. Figure 3.5 shows an example boxplot. Notice that the rich graphical structure means it is quite straightforward to compare two histograms.

3.3 WHOSE IS BIGGER? INVESTIGATING AUSTRALIAN PIZZAS

At `http://www.amstat.org/publications/jse/jse_data_archive.htm`), you will find a dataset giving the diameter of pizzas, measured in Australia (search for the word "pizza"). This website also gives the backstory for this dataset. Apparently, EagleBoys pizza claims that their pizzas are always bigger than Dominos pizzas, and published a set of measurements to support this claim (the measurements were available at `http://www.eagleboys.com.au/realsizepizza` as of Feb 2012, but seem not to be there anymore).

Whose pizzas are bigger? and why? A histogram of all the pizza sizes appears in figure 3.6. We would not expect every pizza produced by a restaurant to have exactly the same diameter, but the diameters are probably pretty close to one another, and pretty close to some standard value. This would suggest that we'd expect to see a histogram which looks like a single, rather narrow, bump about a mean. This is not what we see in figure 3.6 — instead, there are two bumps, which suggests two populations of pizzas. This isn't particularly surprising, because we know that some pizzas come from EagleBoys and some from Dominos.

If you look more closely at the data in the dataset, you will notice that each data item is tagged with the company it comes from. We can now easily plot conditional histograms, conditioning on the company that the pizza came from. These appear in figure 3.7. Notice that EagleBoys pizzas seem to follow the pattern we expect — the diameters are clustered tightly around one value — but Dominos pizzas do not seem to be like that. This is reflected in a boxplot (figure 3.8), which shows the range of Dominos pizza sizes is surprisingly large, and that EagleBoys pizza sizes have several large outliers. There is more to understand about this data. The dataset contains labels for the type of crust and the type of topping — perhaps

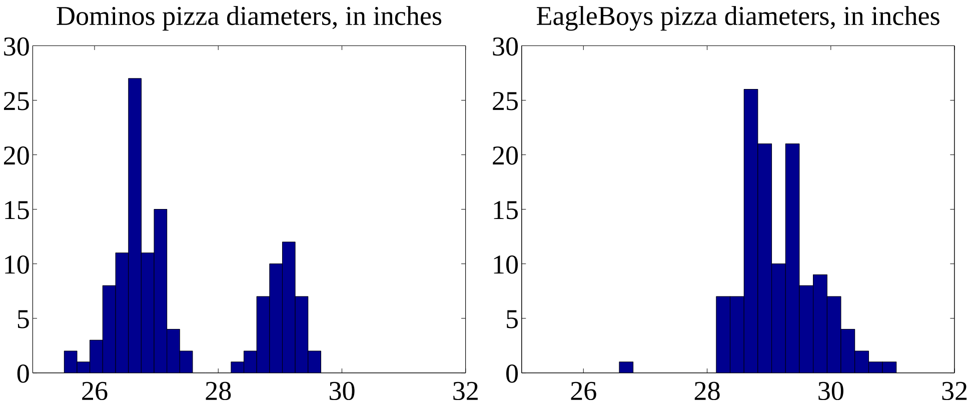

FIGURE 3.7: *On the* **left**, *the class-conditional histogram of Dominos pizza diameters from the pizza data set; on the* **right**, *the class-conditional histogram of EagleBoys pizza diameters. Notice that EagleBoys pizzas seem to follow the pattern we expect — the diameters are clustered tightly around a mean, and there is a small standard deviation — but Dominos pizzas do not seem to be like that. There is more to understand about this data.*

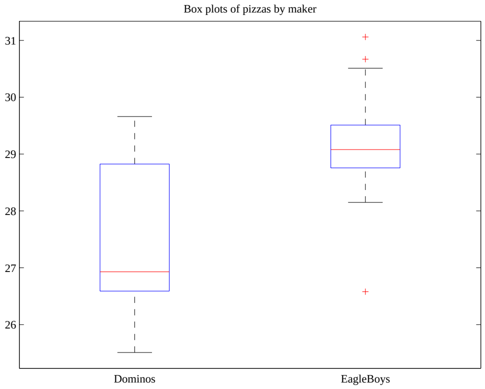

FIGURE 3.8: *Boxplots of the pizza data, comparing EagleBoys and Dominos pizza. There are several curiosities here: why is the range for Dominos so large (25.5-29)? EagleBoys has a smaller range, but has several substantial outliers; why? One would expect pizza manufacturers to try and control diameter fairly closely, because pizzas that are too small present risks (annoying customers; publicity; hostile advertising) and pizzas that are too large should affect profits.*

these properties affect the size of the pizza?

EagleBoys produces DeepPan, MidCrust and ThinCrust pizzas, and Dominos produces DeepPan, ClassicCrust and ThinNCrispy pizzas. This may have something to do with the observed patterns, but comparing six histograms by eye is unattractive. A boxplot is the right way to compare these cases (figure 3.9). The

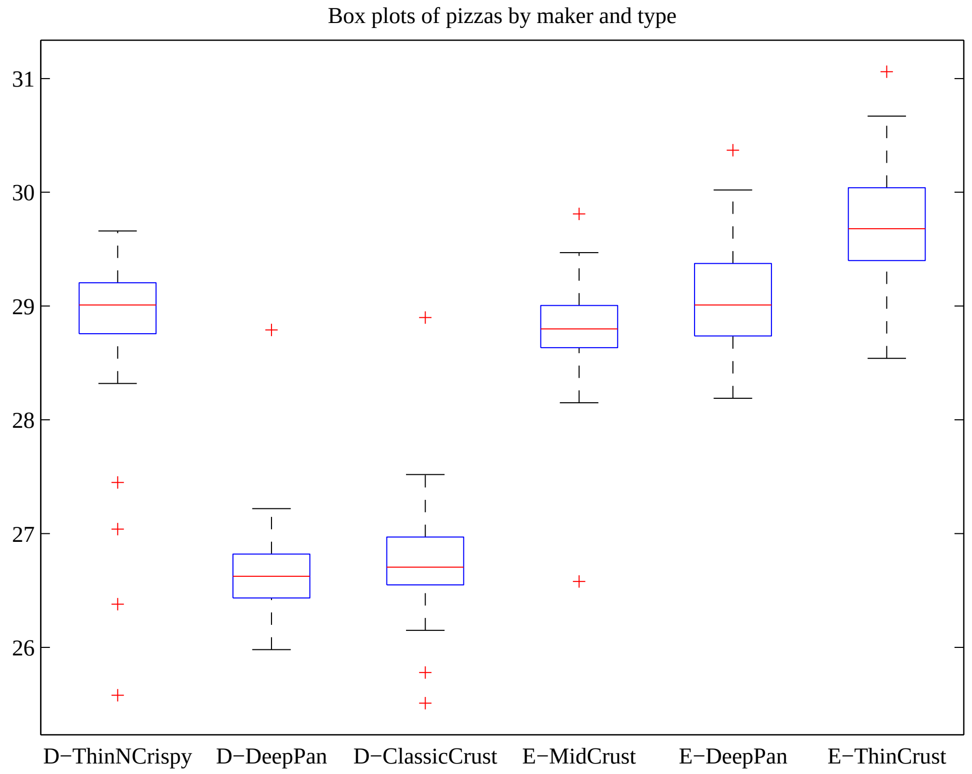

FIGURE 3.9: *Boxplots for the pizza data, broken out by type (thin crust, etc.).*

boxplot gives some more insight into the data. Dominos thin crust appear to have a narrow range of diameters (with several outliers), where the median pizza is rather larger than either the deep pan or the classic crust pizza. EagleBoys pizzas all have a range of diameters that is (a) rather similar across the types and (b) rather a lot like the Dominos thin crust. There are outliers, but few for each type.

Another possibility is that the variation in size is explained by the topping. We can compare types and toppings by producing a set of conditional boxplots (i.e. the diameters for each type and each topping). This leads to rather a lot of boxes (figure 3.10), but they're still easy to compare by eye. The main difficulty is that the labels on the plot have to be shortened. I made labels using the first letter from the manufacturer ("D" or "E"); the first letter from the crust type (previous paragraph); and the first and last letter of the topping. Toppings for Dominos are: Hawaiian; Supreme; BBQMeatlovers. For EagleBoys, toppings are: Hawaiian; SuperSupremo; and BBQMeatlovers. This gives the labels: 'DCBs'; (Dominos; ClassicCrust; BBQMeatlovers); 'DCHn'; 'DCSe'; 'DDBs'; 'DDHn'; 'DDSe'; 'DTBs'; 'DTHn'; 'DTSe'; 'EDBs'; 'EDHn'; 'EDSo'; 'EMBs'; 'EMHn'; 'EMSo'; 'ETBs'; 'ETHn'; 'ETSo'. Figure 3.10 suggests that the topping isn't what is important, but the crust (group the boxplots by eye).

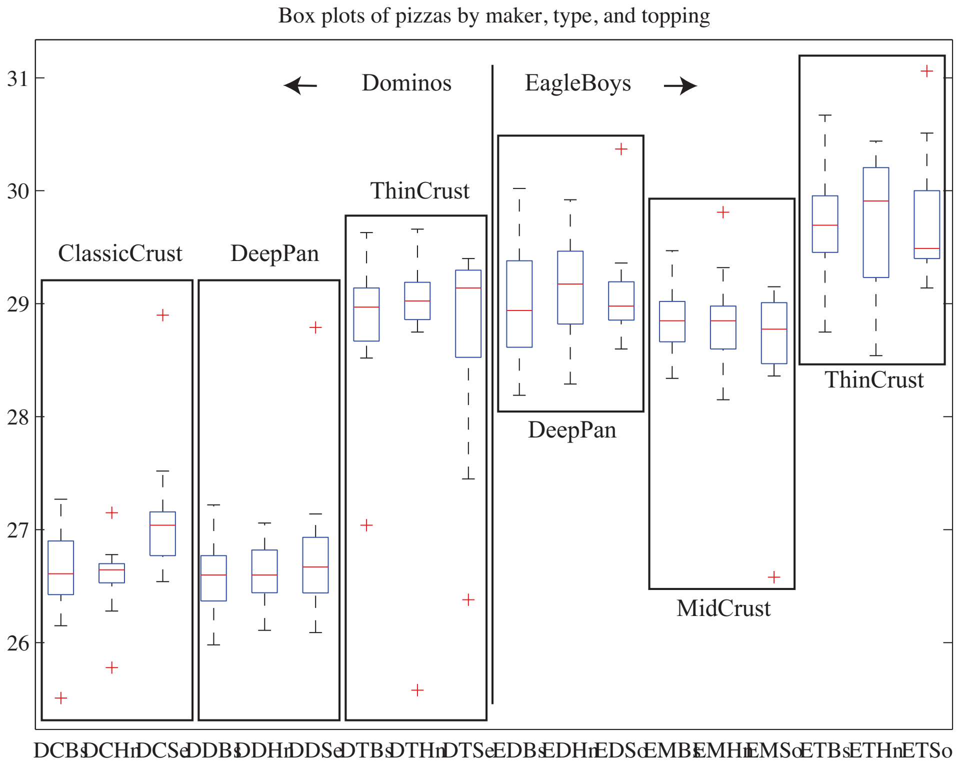

FIGURE 3.10: *The pizzas are now broken up by topping as well as crust type (look at the source for the meaning of the names). I have separated Dominos from Eagleboys with a vertical line, and grouped each crust type with a box. It looks as though the issue is not the type of topping, but the crust. Eagleboys seems to have tighter control over the size of the final pizza.*

What could be going on here? One possible explanation is that Eagleboys have tighter control over the size of the final pizza. One way this could happen is that all EagleBoys pizzas start the same size and shrink the same amount in baking, whereas all Dominos pizzas start a standard diameter, but different Dominos crusts shrink differently in baking. Another way is that Dominos makes different size crusts for different types, but that the cooks sometimes get confused. Yet another possibility is that Dominos controls portions by the mass of dough (so thin crust diameters tend to be larger), but Eagleboys controls by the diameter of the crust.

You should notice that this is more than just a fun story. If you were a manager at a pizza firm, you'd need to make choices about how to control costs. Labor costs, rent, and portion control (i.e. how much pizza, topping, etc. a customer gets for their money) are the main thing to worry about. If the same kind of pizza has a wide range of diameters, you have a problem, because some customers are getting too much (which affects your profit) or too little (which means they might call someone else). But making more regular pizzas might require more skilled (and so more expensive) labor. The fact that Dominos and EagleBoys seem to be following different strategies successfully suggests that more than one strategy might work. But you can't choose if you don't know what's happening. As I said at the start, "what's going on here?" is perhaps the single most useful question anyone can ask.

3.4 NORMALIZED 2D SCATTER PLOTS

As you recall from section 2.3.4, scale is a problem for scatter plots. The way to avoid the problem is to plot in standard coordinates.

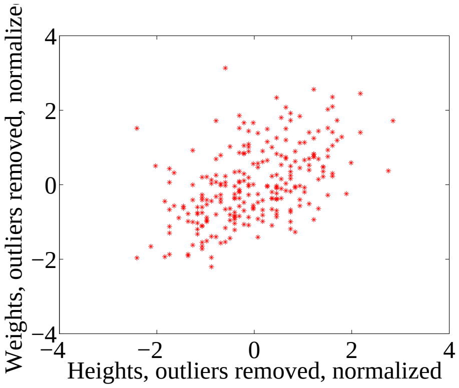

FIGURE 3.11: *A normalized scatter plot of weight against height, from the dataset at* `http://www2.stetson.edu/~jrasp/data.htm`*. Now you can see that someone who is a standard deviation taller than the mean will tend to be somewhat heavier than the mean too.*

A natural solution to problems with scale is to normalize the x and y coordinates of the two-dimensional data to 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. 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 $\hat{x}_j = (x_j - \mathsf{mean}\left(\{x\}\right))/\mathsf{std}\left(x\right))$ 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.

3.5 CORRELATION

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 $\hat{x}$

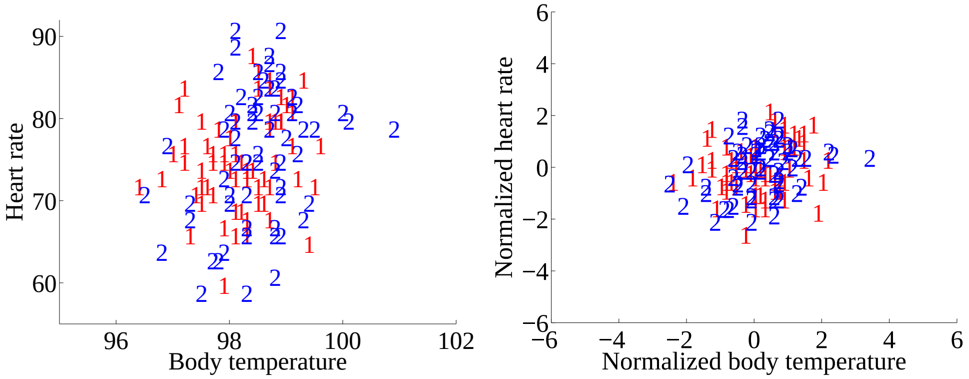

FIGURE 3.12: **Left:** *A scatter plot of body temperature against heart rate, from the dataset at* `http://www2.stetson.edu/~jrasp/data.htm`*; normtemp.xls. I have separated the two genders by plotting a different symbol for each (though I don't know which gender is indicated by which letter); if you 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.*

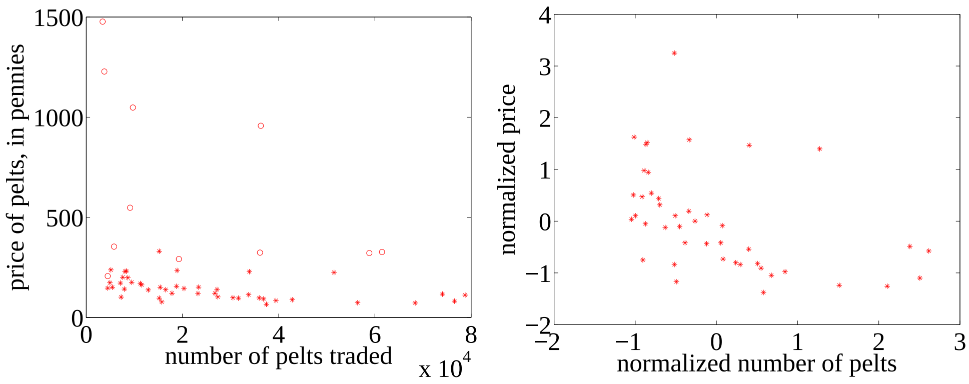

FIGURE 3.13: **Left:** *A scatter plot of the price of lynx pelts against the number of pelts (this is a repeat of figure 2.13 for reference). I have plotted data for 1901 to the end of the series as circles, and the rest of the data as *'s. It is quite hard to draw any conclusion from this data, because 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.*

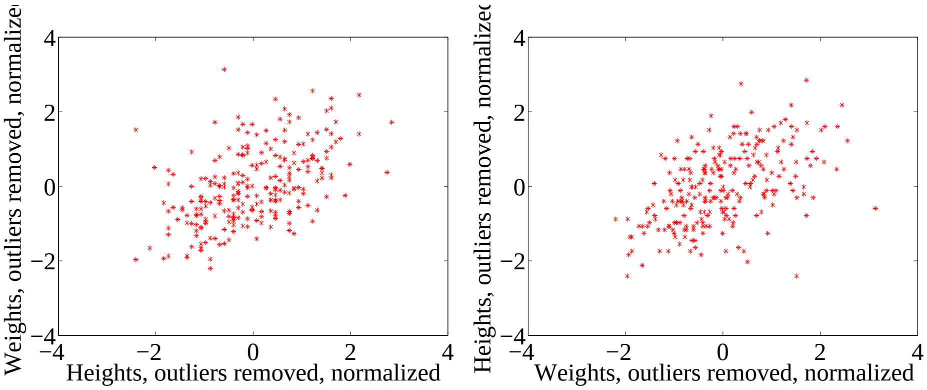

FIGURE 3.14: *On the* **left**, *a normalized scatter plot of weight (y-coordinate) against height (x-coordinate). On the* **right**, *a scatter plot of height (y-coordinate) against weight (x-coordinate). I've put these plots next to one another so you don't have to mentally rotate (which is what you should usually do).*

values tend to appear with smaller $\hat{y}$ values. I have shown these cases together in one figure using a real data example (Figure 3.15), so you can compare the appearance of the plots.

Positive correlation occurs when larger $\hat{x}$ values tend to appear with larger $\hat{y}$ values. This means that data points with 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 3.11 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.

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 $\mathsf{mean}\left(\{\hat{y}\}\right) = 0$, and $\mathsf{var}\left(\{\hat{y}\}\right) = 1$. Our value of $\hat{y}$ should have this mean and this variance, but it doesn't depend on $\hat{x}$ in any way. This is enough information to predict what the plot will look like. We know that $\mathsf{mean}\left(\{\hat{x}\}\right) = 0$ and $\mathsf{var}\left(\{\hat{x}\}\right) = 1$; 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,

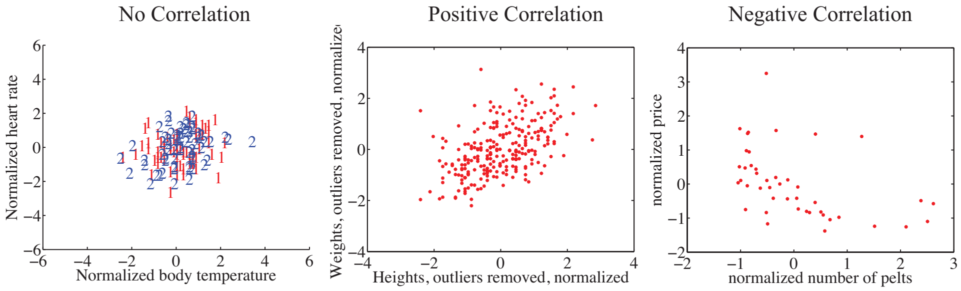

FIGURE 3.15: *The three kinds of scatter plot are less clean for real data than for our idealized examples. Here I used the body temperature vs heart rate data for the zero correlation; the height-weight data for positive correlation; and the lynx data for negative correlation. The pictures aren't idealized — real data tends to be messy — but you can still see the basic structures.*

and then seldom. So for a strip with $\hat{x}$ close to zero, we might see large $\hat{y}$ values; but for one that is far away, we expect to see small $\hat{y}$ values. We should see a blob, centered at the origin. In the temperature-heart rate plot of figure 3.12, 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.

Negative correlation occurs when larger $\hat{x}$ values tend to appear with smaller $\hat{y}$ values. This means that data points with 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 3.13 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.

Notice that leaving out data, as I did here, should be done 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.

The correlation is not affected by which variable is plotted on the x-axis and which is plotted on the y-axis. Figure 3.14 compares a plot of height against weight to one of weight against height. Usually, one just does this by rotating the page, 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.

3.5.1 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 $\mathsf{mean}\left(\{\hat{x}\}\right) = 0$; $\mathsf{mean}\left(\{\hat{y}\}\right) = 0$; $\mathsf{std}\left(\hat{x}\right) = 1$; and $\mathsf{std}\left(\hat{y}\right) = 1$. Each of these summaries is itself the mean of some monomial. So $\mathsf{std}\left(\hat{x}\right)^2 = \mathsf{mean}\left(\left\{\hat{x}^2\right\}\right) = 1$; $\mathsf{std}\left(\hat{y}\right)^2 = \mathsf{mean}\left(\left\{\hat{y}^2\right\}\right)$ (the other two are easy). We can rewrite this information in terms of means of monomials, giving $\mathsf{mean}\left(\{\hat{x}\}\right) = 0$; $\mathsf{mean}\left(\{\hat{y}\}\right) = 0$; $\mathsf{mean}\left(\left\{\hat{x}^2\right\}\right) = 1$; and $\mathsf{mean}\left(\left\{\hat{y}^2\right\}\right) = 1$. There is one monomial missing here, which is $\hat{x}\hat{y}$.

The term $\mathsf{mean}\left(\{\hat{x}\hat{y}\}\right)$ captures correlation between x and y. The term is known as the **correlation coefficient** or **correlation**.

Definition: 3.11 *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 $\hat{x}_i = \frac{(x_i - \mathsf{mean}(\{x\}))}{\mathsf{std}(x)}$, $\hat{y}_i = \frac{(y_i - \mathsf{mean}(\{y\}))}{\mathsf{std}(y)}$. The correlation coefficient is the mean value of $\hat{x}\hat{y}$, and can be computed as:

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

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 3.16 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.

It's not always convenient or a good idea to produce scatter plots in standard

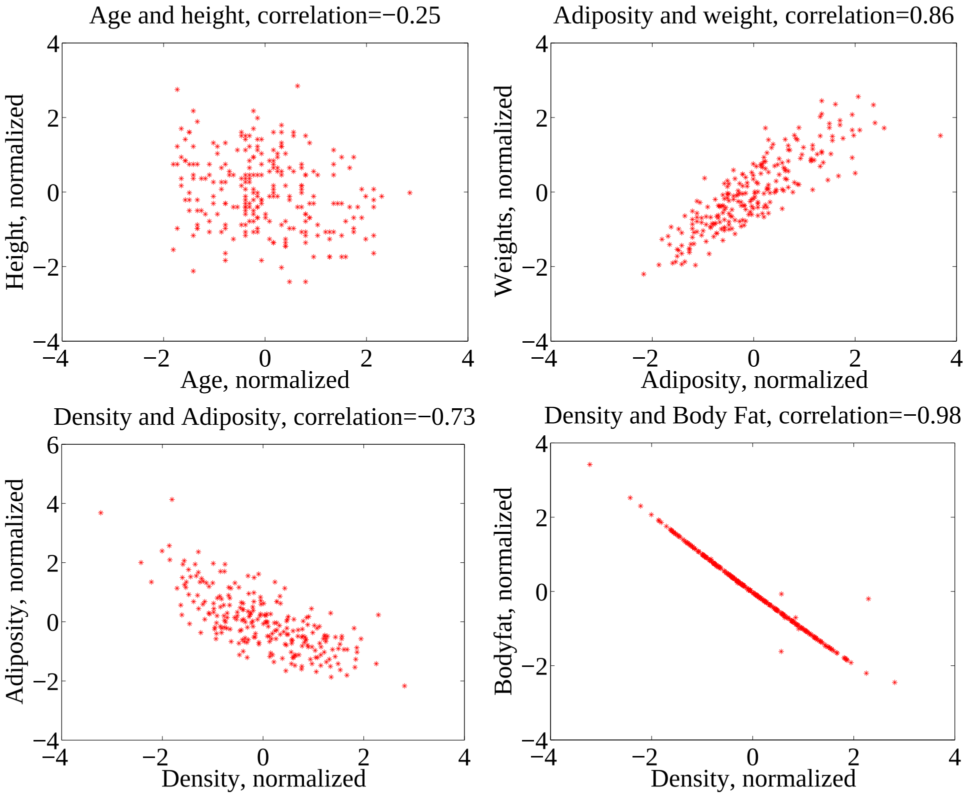

FIGURE 3.16: *Scatter plots for various pairs of variables for the age-height-weight dataset from* `http://www2.stetson.edu/~jrasp/data.htm`*; bodyfat.xls. In each case, two outliers have been removed, and the plots are in standard coordinates (compare to figure 3.17, which shows these data sets plotted in their original units). The legend names the variables.*

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 3.17 shows different correlated datasets plotted in their original units. These data sets are the same as those used in figure 3.16

Properties of the Correlation Coefficient

You should memorize the following properties of the correlation coefficient:

- The correlation coefficient is symmetric (it doesn't depend on the order of its arguments), so
$$\mathsf{corr}\left(\{(x,y)\}\right) = \mathsf{corr}\left(\{(y,x)\}\right)$$

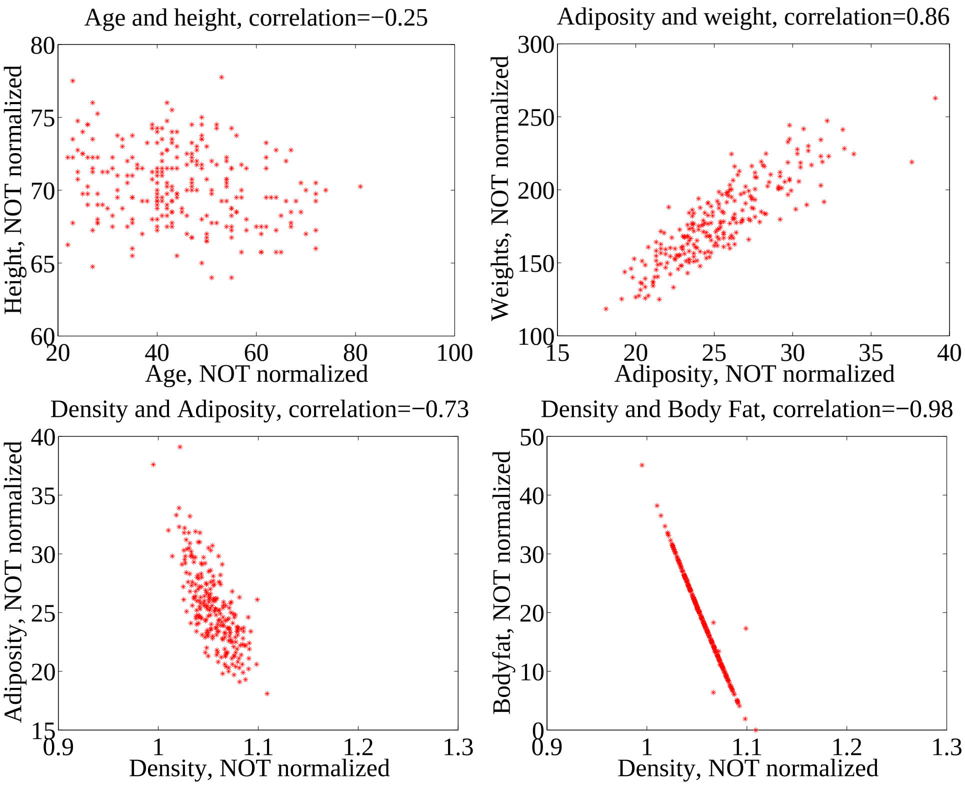

FIGURE 3.17: *Scatter plots for various pairs of variables for the age-height-weight dataset from* `http://www2.stetson.edu/~jrasp/data.htm`*; bodyfat.xls. In each case, two outliers have been removed, and the plots are* NOT *in standard coordinates (compare to figure 3.16, which shows these data sets plotted in normalized coordinates). The legend names the variables.*

- 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

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

- 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 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.

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 $\overline{\hat{y}}(s)$ 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

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

(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 figure 3.15). Then each value of $\overline{\hat{y}}(s)$ is the same — we write $\overline{\hat{y}}$ — and we can rearrange to get

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

Now notice that

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

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

Proposition:

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

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.

Property 3.4: 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.

3.5.2 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

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

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

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

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 $\hat{y}_i$ is NOT $1/r$; you should work this example, which appears in the exercises. We now have a prediction procedure, outlined below.

Procedure: 3.1 *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

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

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

Procedure: 3.2 *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: 3.3 *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

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

so the root mean square error will be $\sqrt{1-r^2}$. This is yet another intepretation 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 figure 2.7, where (before 19h00), prices go down when the number of pelts goes up, and vice versa. These two variables are negatively correlated.

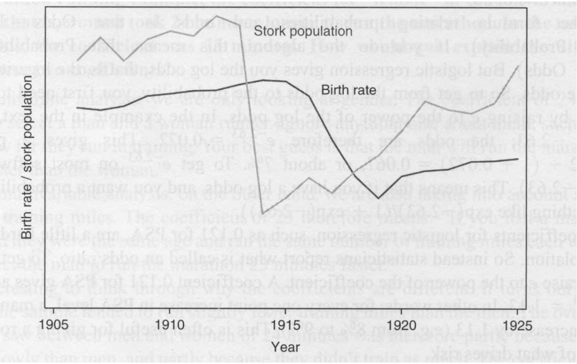

FIGURE 3.18: *This figure, from Vickers (ibid, p184) shows a plot of the stork population as a function of time, and the human birth rate as a function of time, for some years in Germany. The correlation is fairly clear; but this does not mean that reducing the number of storks means there are fewer able to bring babies. Instead, this is the impact of the first world war — a hidden or latent variable.*

3.5.3 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 vectors — there is a pretty good chance that there is some correlation between the coefficients. 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 example illustrates why you need to take great care in setting up experiments and interpreting their results.

This sort of thing happens often, and it's an effect you should look for. Another nice example comes from Vickers (ibid). The graph, shown in Figure 3.18, shows a plot of (a) a dataset of the stork population in Europe over a period of years and (b) a dataset of the birth rate over those years. This isn't a scatter plot; instead, the data has been plotted on a graph. You can see by eye that these two datasets are quite strongly correlated . Even more disturbing, the stork population dropped somewhat before the birth rate dropped. Is this evidence that storks brought babies in Europe during those years? No (the usual arrangement seems to have applied). For a more sensible explanation, look at the dates. The war disturbed both stork and human breeding arrangements. Storks were disturbed immediately by bombs, etc., and the human birth rate dropped because men died at the front.

3.6 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 figure 3.19. 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

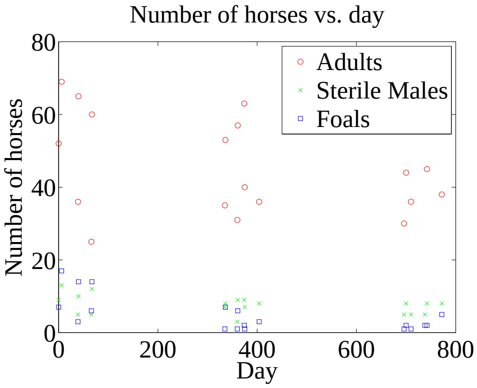

FIGURE 3.19: *A plot of the number of adult horses, sterile males, and foals in horse herds over a period of three years. The plot suggests that introducing sterile males might cause the number of foals to go down. Data from* ***http://lib.stat.cmu.edu/DASL/Datafiles/WildHorses.html.***

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 figure 3.20 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 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 figure 3.20 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 figure 3.21. 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 figure 3.19 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 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

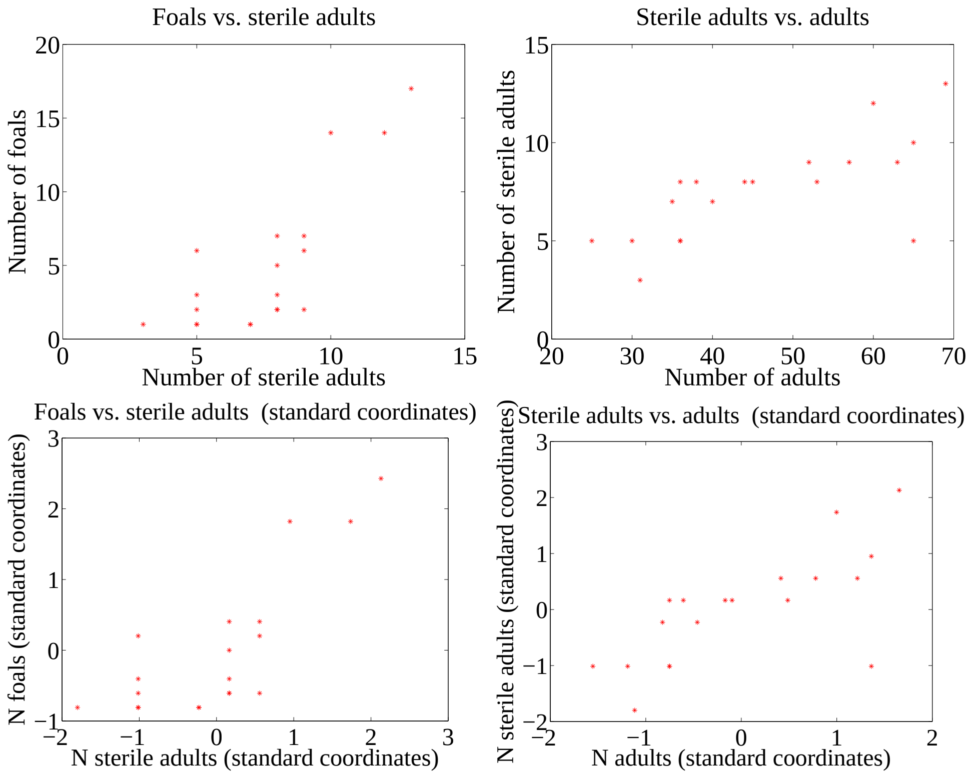

FIGURE 3.20: *Scatter plots of the number of sterile males in a horse herd against the number of adults, and the number of foals against the number of sterile males, from data of* `http://lib.stat.cmu.edu/DASL/Datafiles/WildHorses.html`. **Top:** *unnormalized;* **bottom:** *standard coordinates.*

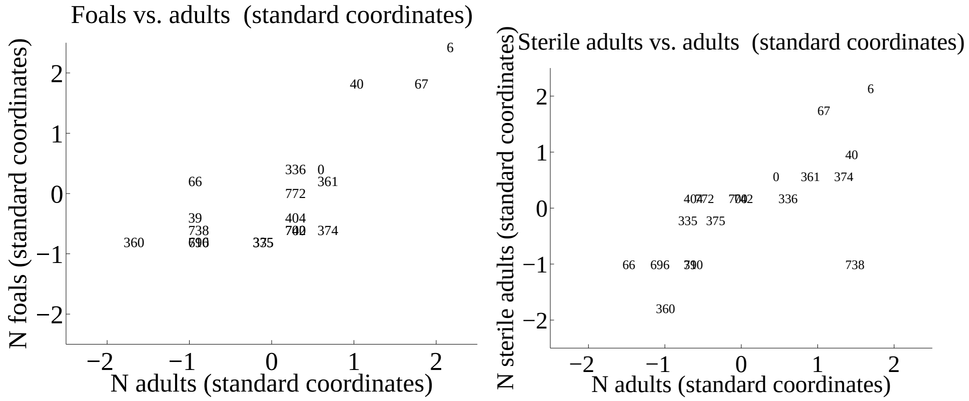

FIGURE 3.21:

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.

3.7 HIGHER DIMENSIONS

We could extract a pair of elements and construct various plots. For vector data, we could also compute the correlation between different pairs of elements. But if each data item is d-dimensional, there could be a lot of pairs to deal with.

We will think of our dataset as a collection of d dimensional vectors. It turns out that there are easy generalizations of our summaries. However, is hard to plot d-dimensional vectors. We need to find some way to make them fit on a 2-dimensional plot. Some simple methods can offer insights, but to really get what is going on we need methods that can at all pairs of relationships in a dataset in one go.

These methods visualize the dataset as a "blob" in a d-dimensional space. Many such blobs are flattened in some directions, because components of the data are strongly correlated. Finding the directions in which the blobs are flat yields methods to compute lower dimensional representations of the dataset.

Now assume that our data items are vectors. This means that we can add and subtract values and multiply values by a scalar without any distress. This is an important assumption, but it doesn't necessarily mean that data is continuous (for example, you can meaningfully add the number of children in one family to the number of children in another family). It does rule out a lot of discrete data. For example, you can't add "sports" to "grades" and expect a sensible answer.

Notation: Our data items are vectors, and we write a vector as $\mathbf{x}$. The data items are d-dimensional, and there are N of them. The entire data set is $\{\mathbf{x}\}$. When we need to refer to the i'th data item, we write $\mathbf{x}_i$. We write $\{\mathbf{x}_i\}$ for a new dataset made up of N items, where the i'th item is $\mathbf{x}_i$. If we need to refer to the j'th component of a vector $\mathbf{x}_i$, we will write $x_i^{(j)}$ (notice this isn't in bold, because it is a component not a vector, and the j is in parentheses because it isn't a power). Vectors are always column vectors.

3.7.1 The Mean

For one-dimensional data, we wrote

$$\mathsf{mean}\left(\{x\}\right) = \frac{\sum_i x_i}{N}.$$

This expression is meaningful for vectors, too, because we can add vectors and divide by scalars. We write

$$\mathsf{mean}\left(\{\mathbf{x}\}\right) = \frac{\sum_i \mathbf{x}_i}{N}$$

and call this the mean of the data. Notice that each component of $\mathsf{mean}\left(\{\mathbf{x}\}\right)$ is the mean of that component of the data. There is not an easy analogue of the median, however (how do you order high dimensional data?) and this is a nuisance. Notice that, just as for the one-dimensional mean, we have

$$\mathsf{mean}\left(\{\mathbf{x} - \mathsf{mean}\left(\{\mathbf{x}\}\right)\}\right) = 0$$

(i.e. if you subtract the mean from a data set, the resulting data set has zero mean).

3.7.2 Parallel Plots

Parallel plots can sometimes reveal information, particularly when the dimension of the dataset is low. To construct a parallel plot, you compute a normalized representation of each component of each data item. The component is normalized by translating and scaling so that the minimum value over the dataset is zero, and the maximum value over the dataset is one. Now write the i'th normalised data item as $(n_1, n_2, \ldots, n_d)$. For this data item, you plot a broken line joining $(1, n_1)$ to $(2, n_2)$ to $(3, n_3$, etc. These plots are superimposed on one another. In the case of the bodyfat dataset, this yields the plot of figure 3.22.

Some structures in the parallel plot are revealing. Outliers often stick out (in figure 3.22, it's pretty clear that there's a data point with a very low height value, and also one with a very large weight value). Outliers affect the scaling, and so make other structures difficult to spot. I have removed them for figure 3.23. In this figure, you can see that two negatively correlated components next to one another produce a butterfly like shape (bodyfat and density). In this plot, you can also see that there are still several data points that are very different from others (two data items have ankle values that are very different from the others, for example).

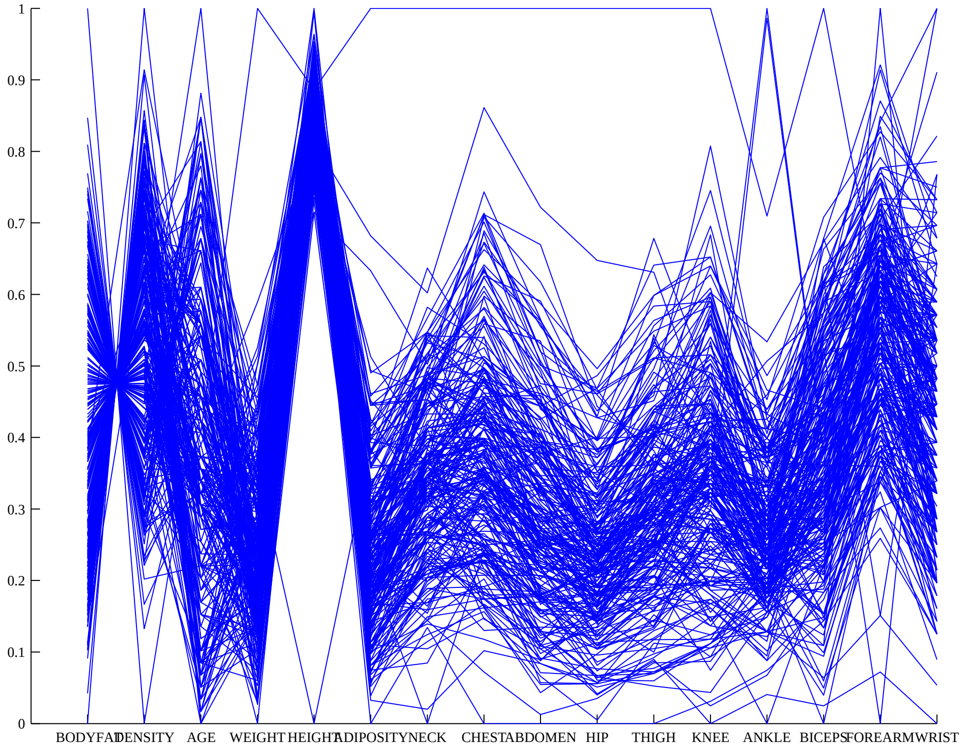

FIGURE 3.22: *A parallel plot of the bodyfat dataset, including all data points. I have named the components on the horizontal axis. It is easy to see that large values of bodyfat correspond to small values of density, and vice versa. Notice that one datapoint has height very different from all others; similarly, one datapoint has weight very different from all others.*

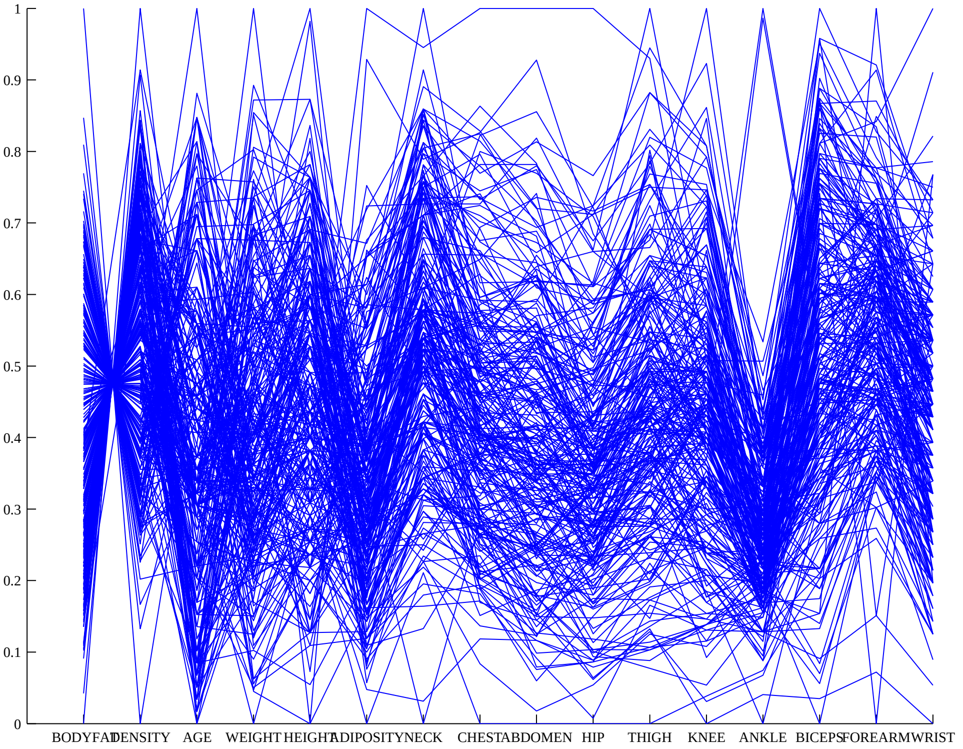

FIGURE 3.23: *A plot with those data items removed, so that those components are renormalized. Two datapoints have rather distinct ankle measurements. Generally, you can see that large knees go with large ankles and large biceps (the v structure).*

C H A P T E R 4

Math Resources

4.1 USEFUL MATERIAL ABOUT MATRICES

Terminology:

- A matrix $\mathcal{M}$ is **symmetric** if $\mathcal{M} = \mathcal{M}^T$. A symmetric matrix is necessarily square.
- We write $\mathcal{I}$ for the identity matrix.
- A matrix is **diagonal** if the only non-zero elements appear on the diagonal. A diagonal matrix is necessarily symmetric.
- A symmetric matrix is **positive semidefinite** if, for any $\mathbf{x}$ such that $\mathbf{x}^T\mathbf{x} > 0$ (i.e. this vector has at least one non-zero component), we have $\mathbf{x}^T\mathcal{M}\mathbf{x} \geq 0$.
- A symmetric matrix is **positive definite** if, for any $\mathbf{x}$ such that $\mathbf{x}^T\mathbf{x} > 0$, we have $\mathbf{x}^T\mathcal{M}\mathbf{x} > 0$.
- A matrix $\mathcal{R}$ is **orthonormal** if $\mathcal{R}^T\mathcal{R} = \mathcal{I} = \mathcal{I}^T = \mathcal{R}\mathcal{R}^T$. Orthonormal matrices are necessarily square.

Orthonormal matrices: You should think of orthonormal matrices as rotations, because they do not change lengths or angles. For $\mathbf{x}$ a vector, $\mathcal{R}$ an orthonormal matrix, and $\mathbf{u} = \mathcal{R}\mathbf{x}$, we have $\mathbf{u}^T\mathbf{u} = \mathbf{x}^T\mathcal{R}^T\mathcal{R}\mathbf{x} = \mathbf{x}^T\mathcal{I}\mathbf{x} = \mathbf{x}^T\mathbf{x}$. This means that $\mathcal{R}$ doesn't change lengths. For $\mathbf{y}$, $\mathbf{z}$ both unit vectors, we have that the cosine of the angle between them is $\mathbf{y}^T\mathbf{x}$; but, by the same argument as above, the inner product of $\mathcal{R}\mathbf{y}$ and $\mathcal{R}\mathbf{x}$ is the same as $\mathbf{y}^T\mathbf{x}$. This means that $\mathcal{R}$ doesn't change angles, either.

Eigenvectors and Eigenvalues: Assume $\mathcal{S}$ is a $d \times d$ symmetric matrix, $\mathbf{v}$ is a $d \times 1$ vector, and λ is a scalar. If we have

$$\mathcal{S}\mathbf{v} = \lambda\mathbf{v}$$

then $\mathbf{v}$ is referred to as an **eigenvector** of $\mathcal{S}$ and λ is the corresponding **eigenvalue**. Matrices don't have to be symmetric to have eigenvectors and eigenvalues, but the symmetric case is the only one of interest to us.

In the case of a symmetric matrix, the eigenvalues are real numbers, and there are d distinct eigenvectors that are normal to one another, and can be scaled to have unit length. They can be stacked into a matrix $\mathcal{U} = [\mathbf{v}_1, \ldots, \mathbf{v}_d]$. This matrix is orthonormal, meaning that $\mathcal{U}^T\mathcal{U} = \mathcal{I}$. This means that there is a diagonal matrix Λ such that

$$\mathcal{S}\mathcal{U} = \mathcal{U}\Lambda.$$

In fact, there is a large number of such matrices, because we can reorder the eigenvectors in the matrix $\mathcal{U}$, and the equation still holds with a new Λ, obtained by reordering the diagonal elements of the original Λ. There is no reason to keep track of this complexity. Instead, we adopt the convention that the elements of $\mathcal{U}$ are always ordered so that the elements of Λ are sorted along the diagonal, with the largest value coming first.

Diagonalizing a symmetric matrix: This gives us a particularly important procedure. We can convert any symmetric matrix $\mathcal{S}$ to a diagonal form by computing

$$\mathcal{U}^T \mathcal{S} \mathcal{U} = \Lambda.$$

This procedure is referred to as **diagonalizing** a matrix. Again, we assume that the elements of $\mathcal{U}$ are always ordered so that the elements of Λ are sorted along the diagonal, with the largest value coming first. Diagonalization allows us to show that positive definiteness is equivalent to having all positive eigenvalues, and positive semidefiniteness is equivalent to having all non-negative eigenvalues.

Factoring a matrix: Assume that $\mathcal{S}$ is symmetric and positive semidefinite. We have that

$$\mathcal{S} = \mathcal{U} \Lambda \mathcal{U}^T$$

and all the diagonal elements of Λ are non-negative. Now construct a diagonal matrix whose diagonal entries are the positive square roots of the diagonal elements of Λ; call this matrix $\Lambda^{(1/2)}$. We have $\Lambda^{(1/2)}\Lambda^{(1/2)} = \Lambda$ and $(\Lambda^{(1/2)})^T = \Lambda^{(1/2)}$. Then we have that

$$\mathcal{S} = (\mathcal{U}\Lambda^{(1/2)})(\Lambda^{(1/2)}\mathcal{U}^T) = (\mathcal{U}\Lambda^{(1/2)})(\mathcal{U}\Lambda^{(1/2)})^T$$

so we can factor $\mathcal{S}$ into the form $\mathcal{X}\mathcal{X}^T$ by computing the eigenvectors and eigenvalues.

4.1.1 Approximating A Symmetric Matrix

Assume we have a $k \times k$ symmetric matrix $\mathcal{T}$, and we wish to construct a matrix $\mathcal{A}$ that approximates it. We require that (a) the rank of $\mathcal{A}$ is precisely $r < k$ and (b) the approximation should minimize the **Frobenius norm**, that is,

$$\| (\mathcal{T} - \mathcal{A}) \|_F{}^2 = \sum_{ij} (T_{ij} - A_{ij})^2.$$

It turns out that there is a straightforward construction that yields $\mathcal{A}$.

The first step is to notice that if $\mathcal{U}$ is orthonormal and $\mathcal{M}$ is any matrix, then

$$\| \mathcal{U}\mathcal{M} \|_F = \| \mathcal{M}\mathcal{U} \|_F = \| \mathcal{M} \|_F.$$

This is true because $\mathcal{U}$ is a rotation (as is $\mathcal{U}^T = \mathcal{U}^{-1}$), and rotations do not change the length of vectors. So, for example, if we write $\mathcal{M}$ as a table of row vectors $\mathcal{M} = [\mathbf{m}_1, \mathbf{m}_2, ...\mathbf{m}_k]$, then $\mathcal{U}\mathcal{M} = [\mathcal{U}\mathbf{m}_1, \mathcal{U}\mathbf{m}_2, ...\mathcal{U}\mathbf{m}_k]$. Now $\| \mathcal{M} \|_F{}^2 = \sum_{j=1}^k \| \mathbf{m}_j \|^2$, so $\| \mathcal{U}\mathcal{M} \|_F{}^2 = \sum_{i=1}^k \| \mathcal{U}\mathbf{m}_k \|^2$. But rotations do not change lengths, so $\| \mathcal{U}\mathbf{m}_k \|^2 = \| \mathbf{m}_k \|^2$, and so $\| \mathcal{U}\mathcal{M} \|_F = \| \mathcal{M} \|_F$. To see the result for the case of $\mathcal{M}\mathcal{U}$, just think of $\mathcal{M}$ as a table of row vectors.

Notice that, if $\mathcal{U}$ is the orthonormal matrix whose columns are eigenvectors of $\mathcal{T}$, then we have

$$\| (\mathcal{T} - \mathcal{A}) \|_F{}^2 = \| \mathcal{U}^T (\mathcal{T} - \mathcal{A}) \mathcal{U} \|_F{}^2.$$

Now write Λ_r for $\mathcal{U}^T \mathcal{A} \mathcal{U}$, and Λ for the diagonal matrix of eigenvalues of $\mathcal{T}$. Then we have

$$\| (\mathcal{T} - \mathcal{A}) \|_F{}^2 = \| \Lambda - \Lambda_A \|_F{}^2,$$

an expression that is easy to solve for Λ_A. We know that Λ is diagonal, so the best Λ_A is diagonal, too. The rank of $\mathcal{A}$ must be r, so the rank of Λ_A must be r as well. To get the best Λ_A, we keep the r largest diagonal values of Λ, and set the rest to zero; Λ_A has rank r because it has only r non-zero entries on the diagonal, and every other entry is zero.

Now to recover $\mathcal{A}$ from Λ_A, we know that $\mathcal{U}^T \mathcal{U} = \mathcal{U} \mathcal{U}^T = \mathcal{I}$ (remember, $\mathcal{I}$ is the identity). We have $\Lambda_A = \mathcal{U}^T \mathcal{A} \mathcal{U}$, so

$$\mathcal{A} = \mathcal{U} \Lambda_A \mathcal{U}^T.$$

We can clean up this representation in a useful way. Notice that only the first r columns of $\mathcal{U}$ (and the corresponding rows of $\mathcal{U}^T$) contribute to $\mathcal{A}$. The remaining $k - r$ are each multiplied by one of the zeros on the diagonal of Λ_A. Remember that, by convention, Λ was sorted so that the diagonal values are in descending order (i.e. the largest value is in the top left corner). We now keep only the top left $r \times r$ block of Λ_A, which we write Λ_r. We then write $\mathcal{U}_r$ for the $k \times r$ matrix consisting of the first r columns of $\mathcal{U}$. Then

$$\mathcal{A} = \mathcal{U}_r \Lambda_r \mathcal{U}^T$$

This is so useful a result, I have displayed it in a box; you should remember it.

Procedure: 4.1 *Approximating a symmetric matrix with a low rank matrix*

Assume we have a symmetric $k \times k$ matrix $\mathcal{T}$. We wish to approximate $\mathcal{T}$ with a matrix $\mathcal{A}$ that has rank $r < k$. Write $\mathcal{U}$ for the matrix whose columns are eigenvectors of $\mathcal{T}$, and Λ for the diagonal matrix of eigenvalues of $\mathcal{A}$ (so $\mathcal{A}\mathcal{U} = \mathcal{U}\Lambda$). Remember that, by convention, Λ was sorted so that the diagonal values are in descending order (i.e. the largest value is in the top left corner).
Now construct Λ_r from Λ by setting the $k - r$ smallest values of Λ to zero, and keeping only the top left $r \times r$ block. Construct $\mathcal{U}_r$, the $k \times r$ matrix consisting of the first r columns of $\mathcal{U}$. Then

$$\mathcal{A} = \mathcal{U}_r \Lambda_r \mathcal{U}_r^T$$

is the best possible rank r approximation to $\mathcal{T}$ in the Frobenius norm.

Now if $\mathcal{A}$ is positive semidefinite (i.e. if at least the r largest eigenvalues of $\mathcal{T}$ are non-negative), then we can factor $\mathcal{A}$ as in the previous section. This yields a procedure to approximate a symmetric matrix by factors. This is so useful a result, I have displayed it in a box; you should remember it.

Procedure: 4.2 *Approximating a symmetric matrix with low dimensional factors*

Assume we have a symmetric $k \times k$ matrix $\mathcal{T}$. We wish to approximate $\mathcal{T}$ with a matrix $\mathcal{A}$ that has rank $r < k$. We assume that at least the r largest eigenvalues of $\mathcal{T}$ are non-negative. Write $\mathcal{U}$ for the matrix whose columns are eigenvectors of $\mathcal{T}$, and Λ for the diagonal matrix of eigenvalues of $\mathcal{A}$ (so $\mathcal{A}\mathcal{U} = \mathcal{U}\Lambda$). Remember that, by convention, Λ was sorted so that the diagonal values are in descending order (i.e. the largest value is in the top left corner).
Now construct Λ_r from Λ by setting the $k - r$ smallest values of Λ to zero and keeping only the top left $r \times r$ block. Construct $\Lambda_r^{(1/2)}$ by replacing each diagonal element of Λ with its positive square root. Construct $\mathcal{U}_r$, the $k \times r$ matrix consisting of the first r columns of $\mathcal{U}$. Then write $\mathcal{V} = (\mathcal{U}_r \Lambda_r^{(1/2)})$

$$\mathcal{A} = \mathcal{V}\mathcal{V}^T$$

is the best possible rank r approximation to $\mathcal{T}$ in the Frobenius norm.