CHAPTER 1

First Tools for Looking at Data

The single most important question for a working scientist is: "what's going on here?" We need to be able to summarize data in useful ways. We need to make pictures of datasets that give a sense of what is going on. This is an activity sometimes known as "Descriptive Statistics". There are a variety of tools we can use.

In this chapter, we will work with 1- and 2- dimensional data, because picturing this data is easier. We look at high dimensional data in chapter 1.3.6.

1.1 SUMMARIES OF 1D DATA

1.1.1 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: *Mean*

Assume we have N data items, $x_1, \ldots, x_N$. Their mean is

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

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 1.1 (you can choose who you want to be in this story). The mean of this data is \$107, 903. One useful way to think of the mean of a dataset is as the best guess of the value of a new data item, given no information at all; 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 $\text{mean}\left(\{kx_i\}\right) = k\text{mean}\left(\{x_i\}\right)$.
- Translating data translates the mean: or $\text{mean}\left(\{x_i + c\}\right) = \text{mean}\left(\{x_i + c\}\right)$.
- The sum of signed differences from the mean is zero. This means that

$$\sum_i (x_i - \text{mean}\left(\{x_i\}\right)) = 0.$$

- Choose the number μ such that the sum of squared distances of data points

Individual	net worth
a	100, 360
b	109, 770
c	96, 860
d	97, 860
e	108, 930
f	124, 330
g	101, 300
h	112, 710
i	106, 740
j	120, 170

TABLE 1.1: *Net worths of people you meet in a bar.*

to μ is minimized. That number is the mean. In notation

$$\arg\min_{\mu} \sum_i (x_i - \mu)^2 = \mathsf{mean}\left(\{x_i\}\right)$$

You can prove the first three by simply writing out the expression; the fourth you prove by taking the derivative and setting to zero.

1.1.2 The Standard Deviation

Definition: *Standard deviation*

Assume we have N data items, $x_1, \ldots, x_N$, where $N > 1$. Their standard deviation is:

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

The standard deviation is the root mean square of the offsets of data from the mean. One should think of the standard deviation as a scale. We will often talk of a data item being "within k standard deviations from the mean" or being "more than k standard deviations from the mean". If the data item is x_i, the mean is μ, and the standard deviation is σ, you should interpret these statements as meaning, respectively

$$\frac{\mathsf{abs}\left(x_i - \mu\right)}{\sigma} \leq k$$

("x_i is within k standard deviations from the mean") and

$$\frac{\mathsf{abs}\left(x_i - \mu\right)}{\sigma} > k$$

("x_i is more than k standard deviations from the mean").

Standard deviation and scale - I Standard deviation has a very important property. For any dataset, it is hard for data items to get many standard deviations away from the mean. Assume we have N data items, x_i. Assume the standard deviation is σ, and the mean is zero (there is no loss of generality here, it just simplifies notation). Then what is the largest fraction r of the data that lies more than k standard deviations away from the mean? To achieve this configuration, our data should have $N(1-r)$ data points with the value 0, and Nr data points with the value $k\sigma$. Because

$$\sigma = \sqrt{\frac{\sum_i x_i^2}{N}}$$

we have that

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

so that

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

So, for example, at most 25% of a dataset is 2 standard deviations away from the mean and at most 1% of a dataset could be 10 standard deviations away from the mean, *for any kind of data at all*. In fact, this bound is wildly pessimistic, because the configuration of data that achieves this bound is very unusual. Most data has more random structure, meaning that we expect to see very much less data 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.

Standard deviation and scale - II Furthermore, *whatever the dataset*, there must be at least one data item that is one or more standard deviations away from the mean. You can see this by looking at the expression for standard deviation. We have

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

Now, this means that

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

But

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

so

$$(\mathsf{sd}\left(\{x_i\}\right))^2 \leq \max_i (x_i - \mathsf{mean}\left(\{x_i\}\right))^2.$$

An example might help. Assume that N is even. Now think about choosing a dataset that has standard deviation one, *and* where all points are as close to the mean as possible. We can do this by placing half the data points at $\mathsf{mean}\left(\{x_i\}\right)+1$ and the other half at $\mathsf{mean}\left(\{x_i\}\right) - 1$. Moving any of these points closer to the mean pushes the standard deviation down; we can only make it bigger by moving some other point further from the mean.

Properties of standard deviation Notice that both mean and standard deviation have the same units as the data. Notice also that for k a constant and $\mathbf{c}$ a constant vector. Standard deviation has two properties you should remember:

1. Scaling data scales the standard deviation: or $\mathsf{sd}\left(\{kx_i\}\right) = k\mathsf{sd}\left(\{x_i\}\right)$
2. Translating data does not change the standard deviation: or $\mathsf{sd}\left(\{x_i + c\}\right) = \mathsf{sd}\left(\{x_i\}\right)$

.

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

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

(notice the $N-1$ for our N). This number is an unbiased estimate of a parameter of a probability distribution that might explain the data. Irritatingly, it 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 - the N in our expressions is the right thing to use for what we're doing.

1.1.3 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 1.1.1. 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 and also 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, it is quite complicated to talk sense about the average net worth of US residents, because the number is strongly affected by age, household income, ethnic origins, and so on. Furthermore, a small number of extremely wealthy people could change the average dramatically, as the example shows. An alternative to using a mean is to use a **median**.

Definition: *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, median([1, 2, 2, 2, 2, 2, 2, 2, 3])=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.

1.1.4 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: *Percentile*

The k'th percentile is the value such that $k\%$ of the data is less than or equal to that value.

Definition: *Quartiles*

The first quartile of the data is the value such that 25% of the data is less than or equal to that value; 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); and the third quartile of the data is the value such that 75% of the data is less than or equal to that value.

Definition: *Interquartile Range*

Write Q_1 for the first quartile, and Q_3 for the third quartile; the interquartile range is $Q_3 - Q_1$.

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.

1.1.5 Calculating summaries

MATLAB is a programming environment widely used in numerical analysis and computer vision circles, among other. We will use Matlab in examples here, and give some guidelines. Many universities, including UIUC, have free student licenses for Matlab, and there are numerous reference books you can look at for details of syntax, etc.

Matlab's focuses on representations of vectors and matrices. For example, we can compute our summaries for our net worth data vector by the matlab code in listing 1.1.

If you want to know what a Matlab command does, you can use `help` or `doc`. For example, you could type `doc prctile`. Generally, I will not put listings in the text unless they make some special point. However, files for code used to make figures is on the website, and I'll refer to those files.

1.1.6 Types of Data

A variable is **continuous** (like, for example, height or weight or body temperature) when you could reasonably expect to encounter any value in a particular range. The number of children a family has is an example of a **categorical** variable — it can take a small number of values (in the US, perhaps 0, 1, 2, . . ., 12), but no others.

1.1.7 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

Listing 1.1: Matlab code used to compute the net worth mean, standard deviation, and interquartile ranges

```
% net worths without billionaire
networths=[100360,109770,96860,97860,...
    108930,124330,101300,112710,...
    106740,120170];
ms=mean(networths); % gives the mean
sds=std(networths); % gives the standard deviation
% net worths with billionaire
bnetworths=[networths, 1e9];
bsms=mean(bnetworths);
bsds=std(bnetworths);
nwpcs=prctile(networths, [25, 50, 75]);
% this gives the specified
% percentiles - first,
% second, and third
% quartiles
nwiqr=nwpcs(3)-nwpcs(1);
% the interquartile range without the billionaire
bnwpcs=prctile(bnetworths, [25, 50, 75]);
bnwiqr=bnwpcs(3)-bnwpcs(1);
% and the interquartile range with the billionaire
```

Andrew Vickers' book *What is a p-value anyway?*), because you can't have fractions of a child. 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. Vickers (ibid) gives the compelling example of a paper reporting the mean length of pregnancy as 32.833 weeks; but 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.

1.2 PLOTTING 1D DATA

1.2.1 Histograms

Assume we have a 1D dataset. This will be a vector of numbers, which are in the same units. For example, each might be a temperature measurement, or a pressure measurement, or a height measurement, or the number of children in a family. A simple, but highly informative, representation is a **histogram**.

Definition: *Histogram*

A histogram represents a dataset by counts of the number of data items in each of a set of boxes. Each data item is counted into exactly one box.

Example: *Histograms in Matlab*

In MATLAB, if you have a vector `x` with one dimensional data in it, you can type `hist(x)` and get a display of a histogram of that data with 10 boxes in the current figure (hint: if you can't see the figure, it might be in a window that is overlapped by the command window). If you do `hist(x, 20)` you can see a histogram with 20 boxes.

Example: *How to build a 1D histogram using even boxes*

Write the i'th number x_i, the smallest value as x_{min} and the largest value as x_{max}. We divide the range between the smallest and largest values into n intervals of even width $(x_{\text{max}} - x_{\text{min}})/n$. The histogram is then an n-dimensional vector of counts; each entry represents the count of the number of data items that lie in that interval. Notice we need to be careful to ensure that each point in the range of values is claimed by exactly one interval. For example, we could have intervals of $[0-1\ )$ and $[1-2\ )$, or we could have intervals of $(0-1\ ]$ and $(1-2\ ]$. We could *not* have intervals of $[0-1]$ and $[1-2]$, because then a data item with the value 1 would appear in two boxes. Similarly, we could not have intervals of $(0-1]$ and $(1-2)$, because then a data item with the value 1 would not appear in any box.

Listing 1.2: Matlab code used to read in pizza data and display a histogram

```
% pizza size data
cd('~/Current/Courses/Probcourse/SomeData/DataSets/');
% this is where I keep the file, but you may have it somewhere else
[num, txt, raw]=xlsread('cleanpizzasize.xls');
sizes=num(:, 5); % these are diameters
figure(1); hist(sizes, 10); figure(1)
% shows a histogram of pizza diameters
```

Example: *How to build a 1D histogram with uneven intervals*

For a histogram with even intervals, it is natural to plot the histogram as a set of boxes, where the height of each box is the number of data items in that box. But a histogram with even intervals can have empty boxes (see figure 1.2). In this case, it can be more informative to have some larger intervals to ensure that each interval has some data items in it. But how high should we plot the box?
Imagine taking two consecutive intervals in a histogram with even intervals, and fusing them. The height of the box over the first interval is n_1, where n_1 is the number of elements in the first box; the height of the box over the second interval is n_2. The interval width is dx, so the area of the box over the first (resp. second) interval is $n_1 dx$ ($n_2 dx$). It is natural that the fused box should have height $(n_1 + n_2)/2$ — the average height of the two boxes. Notice that, if it does, the *area* of that box is $(n_1 + n_2)dx$. For each of the first, second and fused boxes, the area of the box is proportional to the number of elements in the box. This suggests the (correct) rule: plot boxes such that the area is proportional to the number of elements.

Histograms give some useful insight into data. The Matlab code in listing 1.2 will plot histograms of the diameter of pizzas, measured in Australia (you can find a version of this dataset, along with a neat backstory, at `http://www.amstat.org/publications/jse/jse_data_archive.htm`).

The resulting histogram appears in figure 1.1. 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 a mean value. The standard deviation should be quite small. 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 1.1 — instead, there are two bumps, which suggests two populations of pizzas.

1.2.2 Conditional Histograms

In fact, there are two populations of pizzas in figure 1.1, as you know if you looked up and read the backstory. The data comes from a statistical analysis of pizza diameters from two companies, EagleBoys and Dominos (both Australian pizza delivery companies). If you look more closely at the data in the dataset, you will

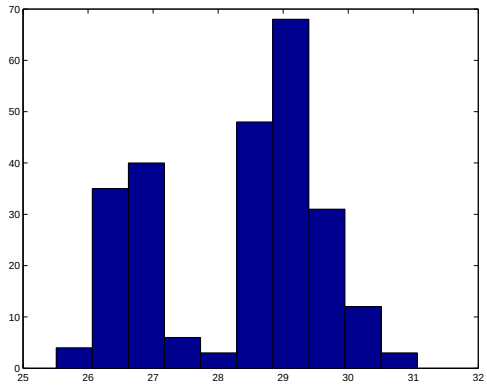

FIGURE 1.1: *A histogram of pizza diameters from the dataset described in the text. Notice that there seem to be two populations.*

Listing 1.3: Matlab code used to normalize axes for the pizza histograms

```
figure(1); axis([25 32 0 70]); figure(1);
figure(2); axis([25 32 0 70]);  figure(2);
figure(3); axis([25 32 0 70]); figure(3);
```

notice that each data item is tagged with the company it comes from. We can now look at the histograms of pizzas from each company separately. These 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 one company). The (rather inelegant) Matlab code in the file `pizzahist2.m` produces two histograms, shown in figure 1.2.

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. You may find figures 1.1 and 1.2 quite difficult to compare, because they are not on the same axes. They can be put on the same axes with appropriate Matlab commands (listing 1.4); the result is figure **??**.

Another interesting dataset is the prices of fish in 1970 and in 1980, which I found at `http://lib.stat.cmu.edu/DASL/Datafiles/FishPrices.html`. One should think about how one looks at this data. If you think about fish as a category, it is reasonable to plot a histogram of the prices. We can then compare the histograms for the two years. I used the Matlab code in the file `fishhists.m`. The first two plots give histograms of the price of fish in 1970 (resp. 1980) plotted in a 5 bin histogram, where Matlab decided on the bin size. But the price of fish went up, so that the 1980 bins are bigger than the 1970 bins, which makes the figures hard to compare (figure 1.4). One can supply Matlab with a vector of bin center points, to get histograms that are a bit easier to compare (Figure 1.5; to learn how to do this, look at `doc hist` or at `fishhists.m`). But if one is interested in *which*

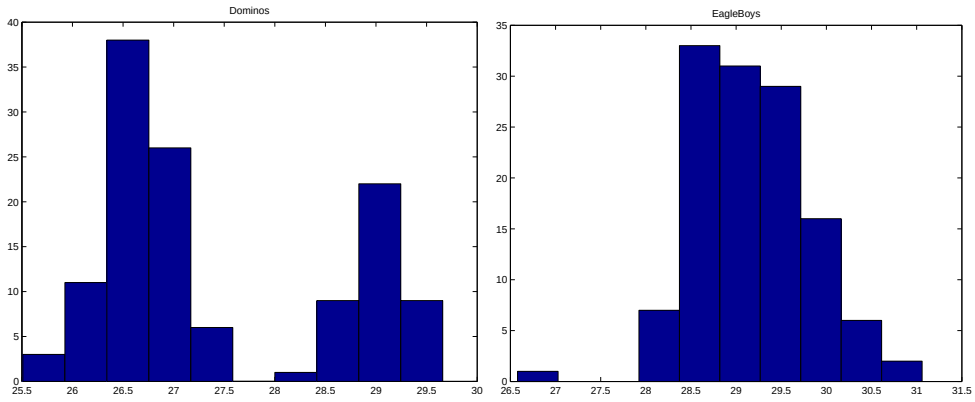

FIGURE 1.2: *On the* **left**, *the class-conditional histogram of Dominos pizza diameters from the pizza data set; on the* **right**, *the class-conditional histogram of EagleBoys pizza diameters. Notice that EagleBoys pizzas seem to follow the pattern we expect — the diameters are clustered tightly 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.*

fish got more expensive, and how fast, a different plot is more helpful (Figure 1.6).

I found data giving the body temperature of a set of individuals at `http://www2.stetson.edu/~jrasp/data.htm`. You might reasonably expect body temperatures to cluster around a small set of numbers and figure 1.7 (a histogram of body temperatures) suggests that they do. I used the matlab code in `temphists.m`. The dataset also gives genders (as 1 or 2 - I don't know which is male and which female). Figure 1.8 gives the class conditional histograms. It does seem like individuals of one gender run a little cooler than individuals of the other. In chapter 1.3.6, we will look at methods to test whether two datasets are different in more detail.

1.2.3 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 the case where there are two peaks (Figure 1.9). 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 1.10). 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 citations to each of the papers they had authored (data

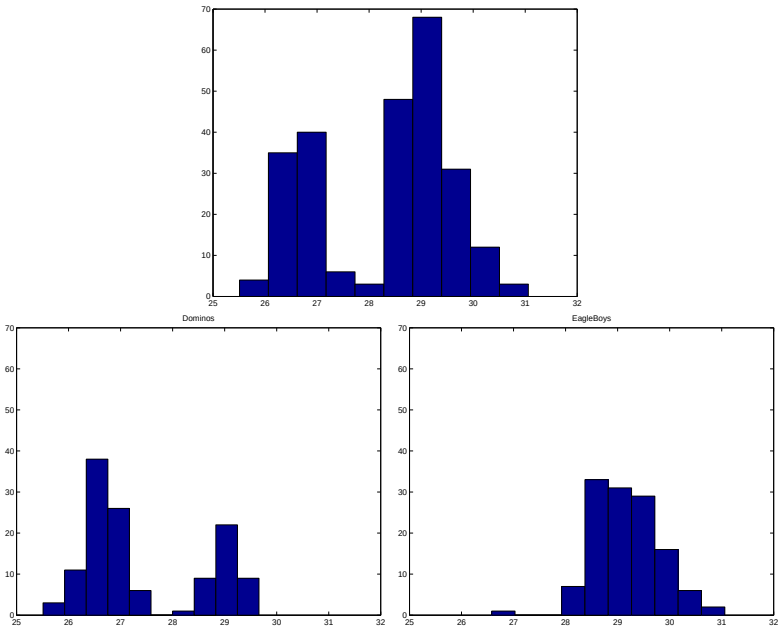

FIGURE 1.3: *The pizza data, now all plotted on the same set of axes.* **Top** *the histogram of all data;* **bottom***, the conditional histograms.*

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 1.11). These are very different from (say) the pizza 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 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 1.12 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 left-skewed, as larger birth weights are somewhat more common than small birth weights.

Skewed data is often, but not always, the result of constraints. For example,

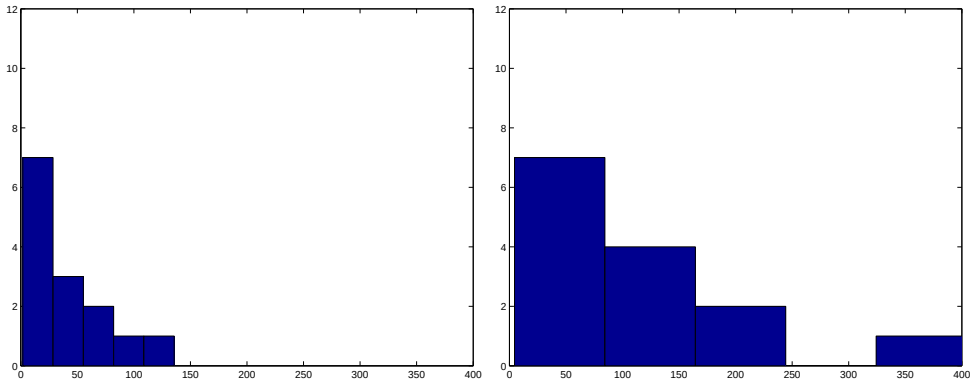

FIGURE 1.4: *Histogram of the price of fish in 1970 (***left***) and 1980 (***right***), each plotted with a 5 bin histogram, where the bins are automatically determined from the data; notice how the bins in the 1980 histogram are bigger, which makes the graphs harder to compare.*

good obstetrical practice means that very large birth weights are extremely rare (birth is typically induced before the baby gets too heavy), but it may be quite hard to avoid some small birth rates. The result could skew birth weights to the left (because large babies will get born, just not as heavy as it might be). 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.

Other things could be going on with the birth weight data, too. Figure 1.13 shows the conditional histograms for birth weights of female and male babies. The histogram for male babies does not appear to be skewed, and there were 26 male babies. The histogram for female babies does look slightly skewed, and there were 18 female babies. It's difficult to be sure if birth weights are skewed from this evidence; perhaps on this day there was an unusual number of light female babies. We need to know more about sampling and probability, and perhaps have more baby data, to come to any conclusion.

1.2.4 Boxplots

Recall that in the pizza dataset, the EagleBoy's pizzas showed the pattern we expected (a narrow group of values around the mean), but the Domino's pizzas did not. The pizza's come in types. 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. We need a different representation of the data.

A **boxplot** is a way to plot data that simplifies comparison. Boxplots show a box for the interquartile range of the data; they show a horizontal line for the median; and they indicate the behavior of the rest of the data with whiskers and/or

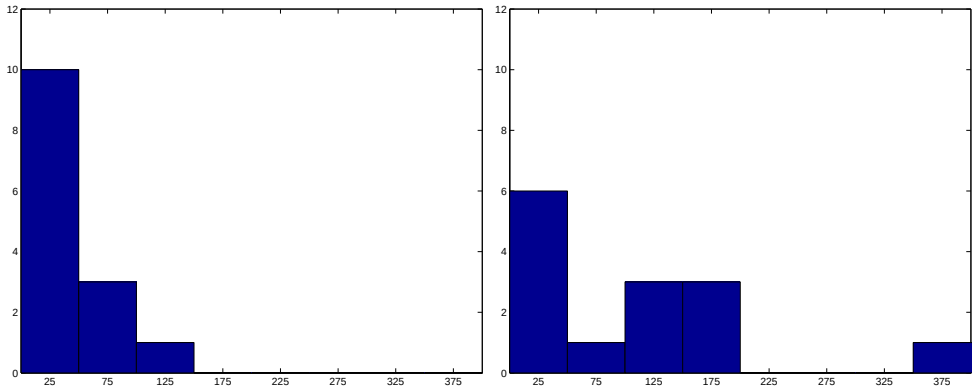

FIGURE 1.5: *Histogram of the price of fish in 1970 (***left***) and 1980 (***right***), each plotted with a histogram where Matlab was given the bin centers. The same set of centers was used for each; now the plots are easier to compare. They're not particularly informative. The main information here is that (a) few kinds of fish are expensive and (b) fish got more expensive. Notice these histograms are certainly not like what one expects from the pizza histograms (a small spread of values symmetrically about a mean); instead, they are left-skewed.*

outliers. The box and median always appear, but there are several choices for the whiskers. Write q_1 for the first quartile and q_3 for the third quartile. Our boxplots assume that any data item larger than $q_3 + 1.5(q_3 - q_1)$ is an outlier, and any data item that is smaller than $q_1 - 1.5(q_3 - q_1)$ is an outlier. The boxplots we show draw whiskers 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. Figure 1.14 shows an example boxplot.

The nice thing about boxplots is that one can compare numerous groups of data quite easily in one plot. I used Matlab to prepare two boxplots. One breaks out the pizzas by type (figure 1.15), the other by type and topping (figure 1.16).

The code for preparing the boxplots is

1.2.5 Standard Coordinates

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**.

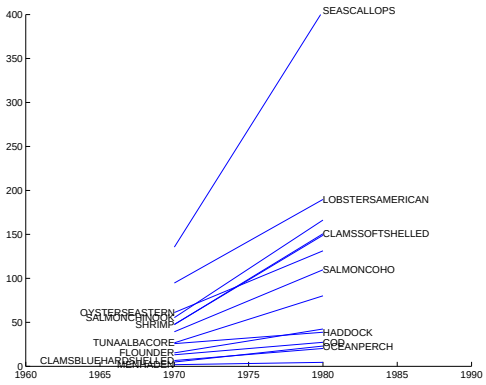

FIGURE 1.6: *An alternative plot of the fish price data. Here we have joined the price in 1970 to the price in 1980 with a line, for each species. This isn't perfect (the layout of the labels is clumsy, and ugly), but it does convey more information than the histograms did. Fish prices went up. Most fish prices went up at about the same rate (most lines are parallel), but some fish prices went up much faster than others (eg SEASCALLOPS).*

Definition: *Standard coordinates*

Assume we have N data items, $x_1, \ldots, x_N$. We represent these data items in standard coordinates by computing

$$\hat{x}_i = \frac{(x_i - \mathsf{mean}\,(\{x_i\}))}{\mathsf{sd}\,(\{x_i\})}.$$

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

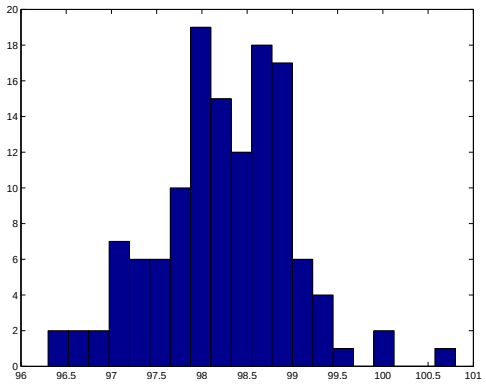

FIGURE 1.7: *Histogram of body temperatures.*

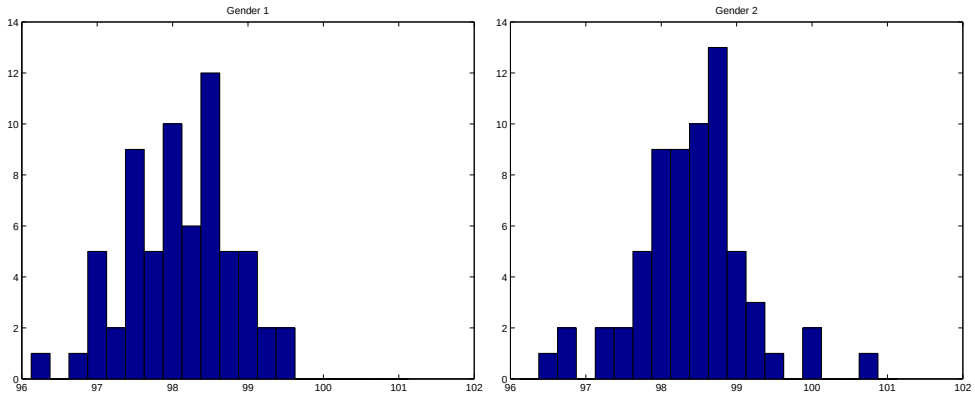

FIGURE 1.8: *Histogram of body temperatures by gender*

coordinates (I sometimes refer to these as "normalized data items"). Then we have

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

Now write $\mu = \mathsf{mean}\left(\{x_i\}\right)$ and $\sigma = \mathsf{sd}\left(\{x_i\}\right)$. Recall that $\mathsf{mean}\left(\{x_i + c\}\right) = \mathsf{mean}\left(\{x_i\}\right) + c$, and that $\mathsf{mean}\left(\{kx_i\}\right) = k\mathsf{mean}\left(\{x_i\}\right)$. We have

$$\begin{aligned}
\mathsf{mean}\left(\{\hat{x}_i\}\right) &= \mathsf{mean}\left(\left\{\frac{(x_i - \mu)}{\sigma}\right\}\right) \\
&= \frac{1}{\sigma}\mathsf{mean}\left(\{(x_i - \mu)\}\right) \\
&= \frac{1}{\sigma}\left(\mathsf{mean}\left(\{x_i\}\right) - \mu\right) \\
&= 0.
\end{aligned}$$

We also have that

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

Recall that $\mathsf{sd}\left(\{x_i + c\}\right) = \mathsf{sd}\left(\{x_i\}\right)$, and that $\mathsf{sd}\left(\{kx_i\}\right) = k\mathsf{sd}\left(\{x_i\}\right)$. We have

$$\begin{aligned}
\mathsf{sd}\left(\{\hat{x}_i\}\right) &= \mathsf{sd}\left(\left\{\frac{(x_i - \mu)}{\sigma}\right\}\right) \\
&= \frac{1}{\sigma}\mathsf{sd}\left(\{(x_i - \mu)\}\right) \\
&= \frac{1}{\sigma}\mathsf{sd}\left(\{x_i\}\right) \\
&= \frac{1}{\sigma}\sigma \\
&= 1.
\end{aligned}$$

An extremely important fact about data is that, for many kinds of data, histograms of these standard coordinates look the same. This is particularly likely for data that results from adding random numbers.

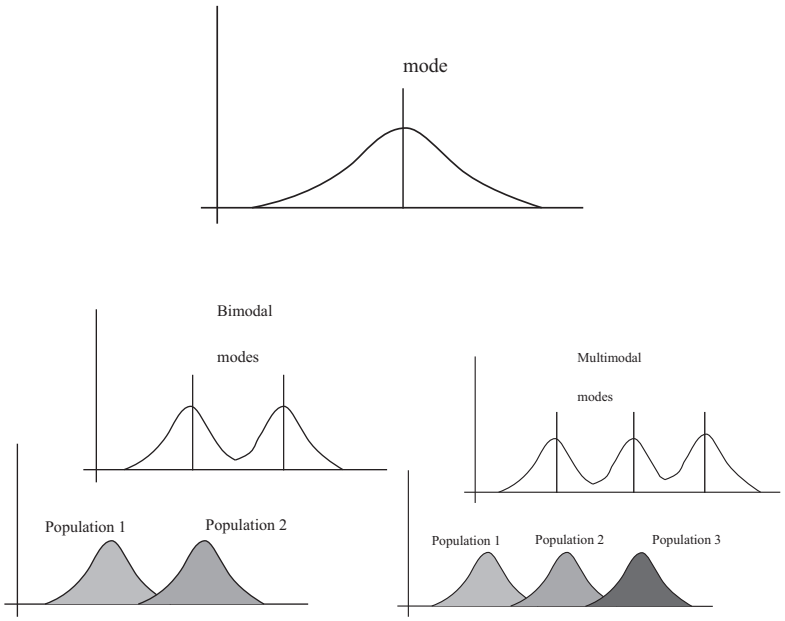

FIGURE 1.9: *Many histograms are unimodal, like the example on the* **top***; there is one peak, or mode. Some are bimodal (two peaks;* **bottom left***) or even multimodal (two or more peaks;* **bottom right***). One common reason (but not the only reason) is that there are actually two populations 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, pyranees mountain dog, etc.).*

1.2.6 Normal Distributions

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 figure 1.17. This phenomenon is so important that data of this form has a special name.

Definition: *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 1.17).

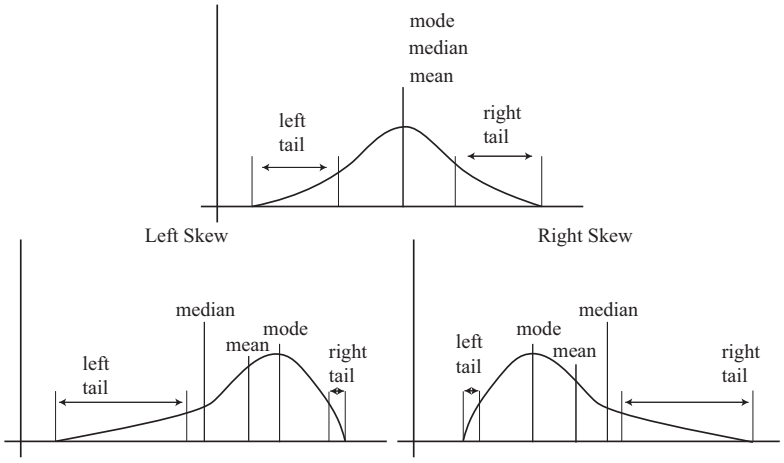

FIGURE 1.10: *On the* **top**, *an example of a symmetric histogram, showing its tails (relatively uncommon values that are significantly larger or smaller than the peak or mode).* **Lower left**, *a sketch of a left-skewed histogram. Here there are few large values, but some very small values that occur 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).*

Definition: *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.

Imagine the following procedure. I flip a coin k times, and record the fraction of heads. I then repeat this experiment N times. Each experiment gives me an estimate of what fraction of coin tosses will come up heads. These fractions are most likely somewhat different, particularly if k is small (if you're worried by this point, think about what happens when $k = 1$ or $k = 2$). We'll build detailed mathematical models of what happens later on, but we can look at data from these experiments now (figure 1.18).

Here are two alternative experiments. In the first, I paint one face of a six-sided die red. I then roll the die k times, and record the fraction of rolls producing a red face. I do this N times to get N numbers. In the second, I paint two faces of a six-sided die red, roll it k times, and record the fraction of rolls resulting in a red face. Again, I do this N times to get N numbers. Figure 1.19 shows the histograms for these experiments. Notice that the histograms do not look the same (the means

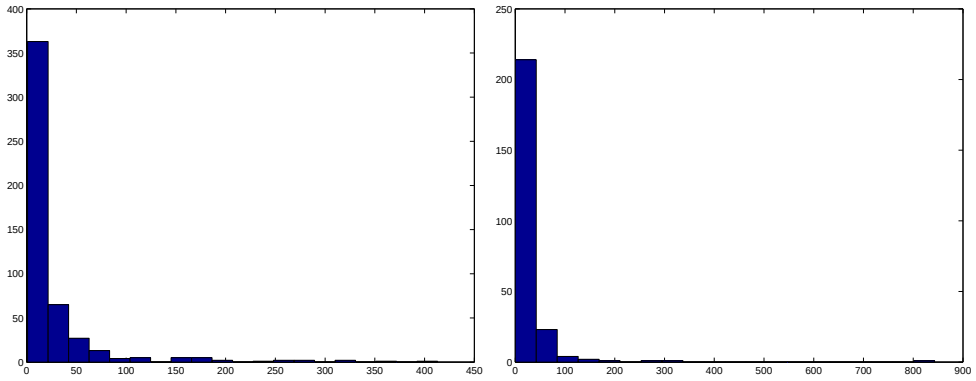

FIGURE 1.11: *Histogram of citations for two faculty members. Notice these histograms are strongly right-skewed.*

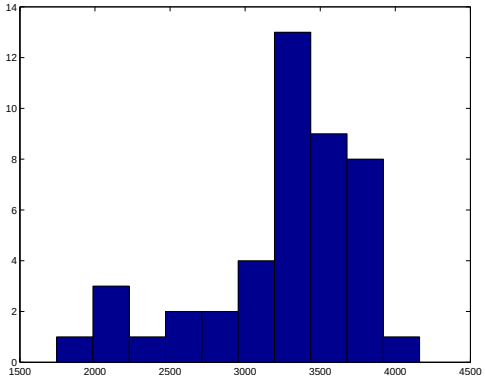

FIGURE 1.12: *Histogram of birth weights for 44 babies borne in Brisbane in 1997. This histogram looks slightly left-skewed.*

are different) but the normalized histograms are very similar.

Figure 1.20 compares normalized histograms for two coin flip experiments and the two die experiments. Notice that they are all very similar. This form is shared by many (but not all) kinds of data; Figure 1.21 shows a variety of different data sets in standard coordinates, each of which looks normal. We will discuss methods to tell whether data is normal later.

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.

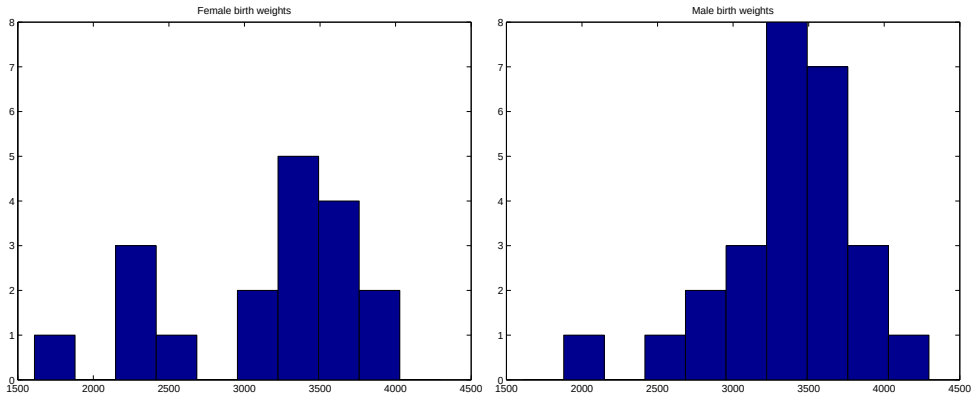

FIGURE 1.13: *Conditional histograms of birth weights for 44 babies born in Brisbane in 1997, by gender. The histogram for male babies does not look skewed, but the histogram for female babies might be. Perhaps on this day there was an unusual number of light babies, or perhaps birth weight is skewed.*

- 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.

1.3 SOME TOOLS FOR 2D DATA

1.3.1 Scatter plots

2D data introduces some new problems. It is not enough to plot a histogram of each coordinate separately, because one coordinate might depend quite strongly on another. One useful representation is a scatter plot. We plot a small shape at the location of each data item $\mathbf{x}_i = (x_i, y_i)$ (Figure 1.22). If there are categorical labels in the data, one can use these to choose a shape. For example, in Figure 1.23, we plot a '1' for data items that have the gender label '1', and a '2' for items that have the gender label '2'. In Figure 1.24, I plot the first seven characters of the state's name.

1.3.2 Correlation

Scatter plots are a useful way to get some sense of what is happening in a dataset. However, the picture in the plot depends on the scale on which one plots the data. For example, plotting lengths in meters gives a very different scatter from plotting lengths in millimeters. Ensuring that all the data appears in the plot can help a

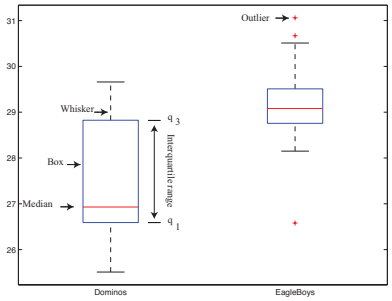

FIGURE 1.14: *A boxplot of pizza data, by manufacturer. The box runs from the first to the third quartile. The horizontal line gives the median. In this boxplot, data items that are larger than $q_3 + 1.5(q_3 - q_1)$ or smaller than $q_1 - 1.5(q_3 - q_1)$, are outliers. The whiskers run from q_1 to the smallest data item that is not an outlier (which is why the lower whisker on the EagleBoys plot is so short) and from q_3 to the largest data item that is not an outlier. These plots are informative. EagleBoys pizzas have quite similar diameters (apart from three outliers). Dominos pizzas have a larger range of diameters, but the median is smaller than the median EagleBoys pizza, and is not in the middle of the range of values. The median is close to q_1. There are no outliers. This means that a quarter of the Dominos pizzas have diameters in the range of the lower whisker; another quarter in the range between q_1 and the median; another quarter in the range between the median and q_3; and the final quarter in the range of the top whisker. Half the diameters are in the range 25.5-27 and the other half is in the range 27-29. 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.*

bit here (because it sets the scale), but this doesn't guarantee a good plot. For example, in Figure 1.22, two outliers set the scale of the plot. Even once these have been removed, we cannot necessarily tell relations between the coordinates.

A natural solution to this problem is to normalize the x and y coordinates of the two-dimensional data to standard coordinates (remember, you compute these by subtracting the mean and dividing by the standard deviation). This shows us the dataset on a standard scale. In Figure 1.25, there are two datasets. In one, there is a relationship between the variables, because the scatter plot in standard coordinates forms a narrow oval. In the other, there is little visible relationship.

Each of these datasets consists of 2D data, and for each of these datasets, we already have some summary information. Because the data is normalized, the mean of the x coordinate is zero and the mean of the y coordinate is zero, too. Furthermore, the standard deviation of the x coordinate is one, and so is the standard deviation of the y coordinate.

Recall from section 1.1.2 that the standard deviation is the square root of the average squared offset from the mean. Assume we have a dataset of N data points.

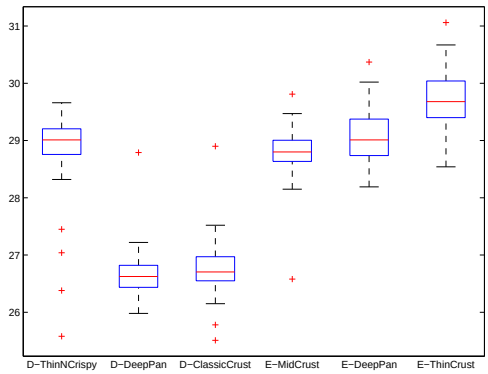

FIGURE 1.15: *Boxplots for the pizza data, broken out by type (thin crust, etc.). This gives some more insight into the data. Dominos thin crust appear to have a narrow range of diameters (with several outliers), where the median pizza is rather larger than either the deep pan or the classic crust pizza. 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. One possible explanation 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 possible explanation is that Dominos makes different size crusts for different types, but that the cooks sometimes get confused.*

We write the i'th datapoint (x_i, y_i). We now normalize this data, and write the i'th data point *in standard coordinates* $(\hat{x}_i, \hat{y}_i)$. The standard deviation of the x coordinate *of the normalized data* can be written as

$$\mathsf{sd}\left(\{x\}\right) = \frac{\sum_i \hat{x}_i^2}{N}.$$

But this is the mean of $\hat{x}^2$ (the square of the x coordinate of the normalized data). We know it to be one (because we normalized the data). In fact, we know $\mathsf{mean}\left(\{\hat{x}_i\}\right)$ (which is 0), $\mathsf{mean}\left(\{\hat{y}_i\}\right)$ (which is 0), $\mathsf{mean}\left(\left\{\hat{x}_i^2\right\}\right)$ (which is 1), $\mathsf{mean}\left(\left\{\hat{y}_i^2\right\}\right)$ (which is 1), and so it is reasonable to wonder about $\mathsf{mean}\left(\{\hat{x}_i\hat{y}_i\}\right)$.

This number is, in fact, a very interesting summary of the data indeed. Imagine we have a data set where there is no relationship between $\hat{x}$ and $\hat{y}$. Then we expect that $\mathsf{mean}\left(\{\hat{x}_i\hat{y}_i\}\right)$ is small or close to zero. One way to see this is to think of a vertical "strip" of the scatter plot (Figure 1.26). If we add up $\hat{x}_j\hat{y}_j$ for j an index choosing only the data points in this strip, the value should look like

$$(\text{center of strip})(\mathsf{mean}\left(\{\hat{y}_j\}\right))$$

but if there is no relationship between $\hat{x}$ and $\hat{y}$, the $\hat{y}$ values should be about the same wherever we put the strip. In turn, this means that $\mathsf{mean}\left(\{\hat{y}_j\}\right)$ is about the same as $\mathsf{mean}\left(\{\hat{y}_i\}\right)$, which is zero. In turn, $\mathsf{mean}\left(\{\hat{x}_i\hat{y}_i\}\right)$ is a sum over strips each of which sums to about zero, so it should be close to zero.

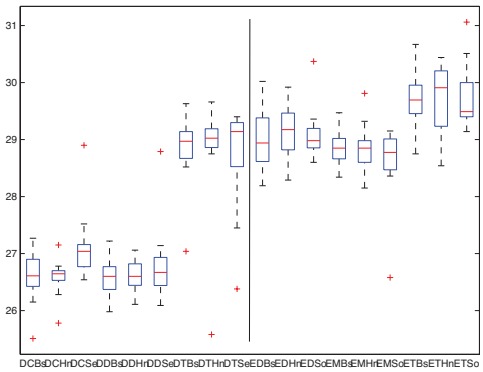

FIGURE 1.16: *The pizzas are now broken up by topping as well as crust type (look at the source for the meaning of the names). To the left of the vertical line is Dominos; to the right is EagleBoys. It looks as though the issue is not the type of topping, but the crust.*

Now imagine if $\hat{y}$ tends to be large when $\hat{x}$ is large, and small when $\hat{x}$ is small. Then the $\hat{y}$'s in a strip will change when we move the strip. Furthermore, the terms in the sum will tend to be positive (because when $\hat{x}$ is positive, so is $\hat{y}$, and when $\hat{x}$ is negative, so is $\hat{y}$). In this case, $\mathsf{mean}\left(\{\hat{x}_i\hat{y}_i\}\right)$ will be positive.

Now imagine if $\hat{y}$ tends to be large when $\hat{x}$ is small, and small when $\hat{x}$ is large. Then the $\hat{y}$'s in a strip will change when we move the strip. Furthermore, the terms in the sum will tend to be negative (if $\hat{x}$ is positive, then $\hat{y}$ tends to be negative, and when $\hat{x}$ is negative, $\hat{y}$ tends to be positive). In this case, $\mathsf{mean}\left(\{\hat{x}_i\hat{y}_i\}\right)$ will be negative.

Finally, it is easy to show that $-1 \leq \mathsf{mean}\left(\{\hat{x}_i\hat{y}_i\}\right) \leq 1$.

This measurement, $\mathsf{mean}\left(\{\hat{x}_i\hat{y}_i\}\right)$ is sufficiently important to have a name; it is known as the **correlation coefficient** or **correlation**.

Definition: *Correlation coefficient*

Assume we have N data items, $(x_1, y_1), \ldots, (x_N, y_N)$, where $N > 1$. Write $\mu_x = \mathsf{mean}\left(\{x_i\}\right)$, $\mu_y = \mathsf{mean}\left(\{y_i\}\right)$, $\sigma_x = \mathsf{sd}\left(\{x_i\}\right)$, $\sigma_y = \mathsf{sd}\left(\{y_i\}\right)$. We compute the correlation coefficient by first normalizing the x and y coordinates to obtain $\hat{x}_i = \frac{(x_i - \mu_x)}{\sigma_x}$, $\hat{y}_i = \frac{(y_i - \mu_y)}{\sigma_y}$. The correlation coefficient is the mean value of $\hat{x}\hat{y}$, and can be computed as:

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

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

$$\mathsf{corr}\left(\{(x_i, y_i)\}\right) = \mathsf{corr}\left(\{(y_i, x_i)\}\right)$$

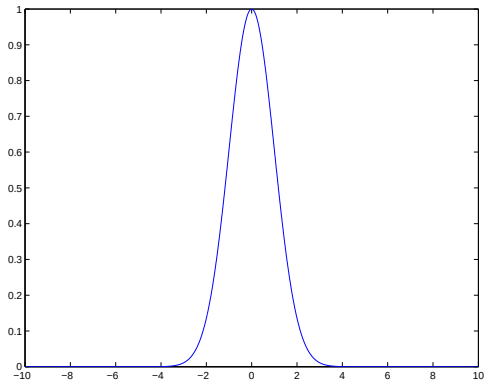

FIGURE 1.17: *Data is standard normal data when its histogram takes a stylized, bell-shaped form, plotted above. One usually requires a lot of data and very small histogram boxes for this form to be reproduced closely. Nonetheless, the histogram for normal data is unimodal (has a single bump) and is symmetric; the tails fall off fairly fast, and there are few data items that are many standard deviations from the mean.*

1.3.3 What Correlation Means

Assume we have a dataset of N points (x_i, y_i). 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. 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.

Recall the argument of Figure 1.26. When the mean of $\hat{y}$ does not depend on $\hat{x}$ (because we get the same pattern of $\hat{y}$'s in any $\hat{x}$ strip), we have that $r = 0$. But in this case, we should predict $\hat{y}_0$ using only information about $\hat{y}$ — $\hat{x}_c$ does not tell us anything. This means the best prediction is $\hat{y}_0 = \mathsf{mean}\left(\{\hat{y}_i\}\right) = 0$. The error of this prediction should be the standard deviation of $\hat{y}$ (which is 1).

Now think about $u_i = \hat{y}_i - r\hat{x}_i$. This has an important interpretation. Assume we predict that at $\hat{x}_i$, $\hat{y}_i$ takes the value $r\hat{x}_i$. The error we make is u_i. We know that $\mathsf{mean}\left(\{u_i\}\right) = \mathsf{mean}\left(\{\hat{y}_i\}\right) - r\mathsf{mean}\left(\{\hat{x}_i\}\right) = 0$ (because everything is in standard coordinates).

We have

$$\begin{aligned}\mathsf{mean}\left(\{\hat{x}_i u_i\}\right) &= \mathsf{mean}\left(\{\hat{x}_i \hat{y}_i\}\right) - r\mathsf{mean}\left(\{\hat{x}_i \hat{x}_i\}\right) \\ &= r - r \\ &= 0.\end{aligned}$$

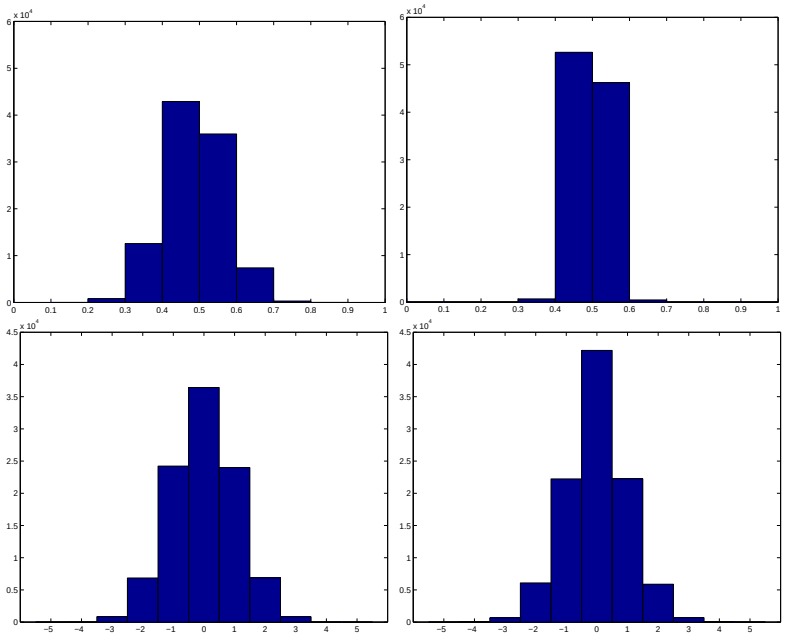

FIGURE 1.18: *In our coin experiment, we flip a coin k times and record the fraction of heads (which has values $0/k, 1/k, \ldots, (k-1)/k, 1$). Each experiment provides one number. We repeat the experiment N times, then plot histograms of the resulting set of numbers.* **Top left:** $k = 40$, $N = 10000$*;* **top right**, $k = 160$, $N = 10000$. *These histograms have been plotted with the same set of boxes. Notice that the spread of data in the case $k = 40$ is considerably greater than in the case $k = 160$. We normalize these datasets by subtracting the location (mean) and dividing by the scale (standard deviation). Each resulting dataset now has mean zero and standard deviation one. We show the histograms for normalized data in for $k = 40$, $N = 10000$ on the* **bottom left** *and for $k = 160$, $N = 10000$ on the* **bottom right**.

In turn, all this means that

$$\mathsf{corr}\left(\{\hat{x}_i, u_i\}\right) = \frac{1}{\mathsf{std}\left(u_i\right)}\mathsf{mean}\left(\{\hat{x}_i u_i\}\right) = 0,$$

i.e. that the correlation between $\hat{x}$ and u is zero. This means that we cannot predict u in any way from $\hat{x}$ — there is no relationship. In turn, this implies our best prediction for $\hat{x}_i$ is to say that $\hat{y}_i$ takes the value $r\hat{x}_i$ — any error that we make by doing so has no relationship to $\hat{x}$.

You can use a version of this argument to establish that if we have the y-coordinate for a data point, but not the x-coordinate, the best prediction for $\hat{x}_i$ (*which is in standard coordinates*) is $r\hat{y}_i$. It is important to notice that the coefficient of $\hat{y}_i$ is NOT $1/r$; you should work this example. This argument gives us a prediction procedure, outlined below.

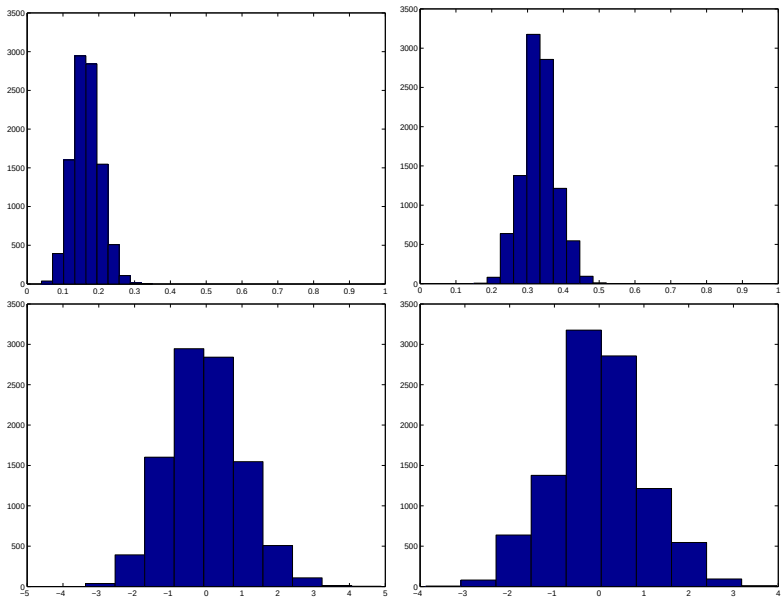

FIGURE 1.19: *In the first die experiment, we take a six-sided die, paint one face red, and roll it k times. We record the fraction of rolls where the red face lands up. We repeat this N times to get N values.* **Top left:** *the histogram of numbers for $k = 100$, $N = 10000$. In the second die experiment, we take a six-sided die, paint two faces red, and roll it k times. We record the fraction of rolls where a red face lands up. We repeat this N times to get N values.* **Top right:** *the histogram of numbers for $k = 100$, $N = 10000$. Once normalised, these histograms again look like one another.* **Bottom left:** *shows the normalized histogram for the one face experiment and* **bottom right** *shows the normalized histogram for the two face experiment.*

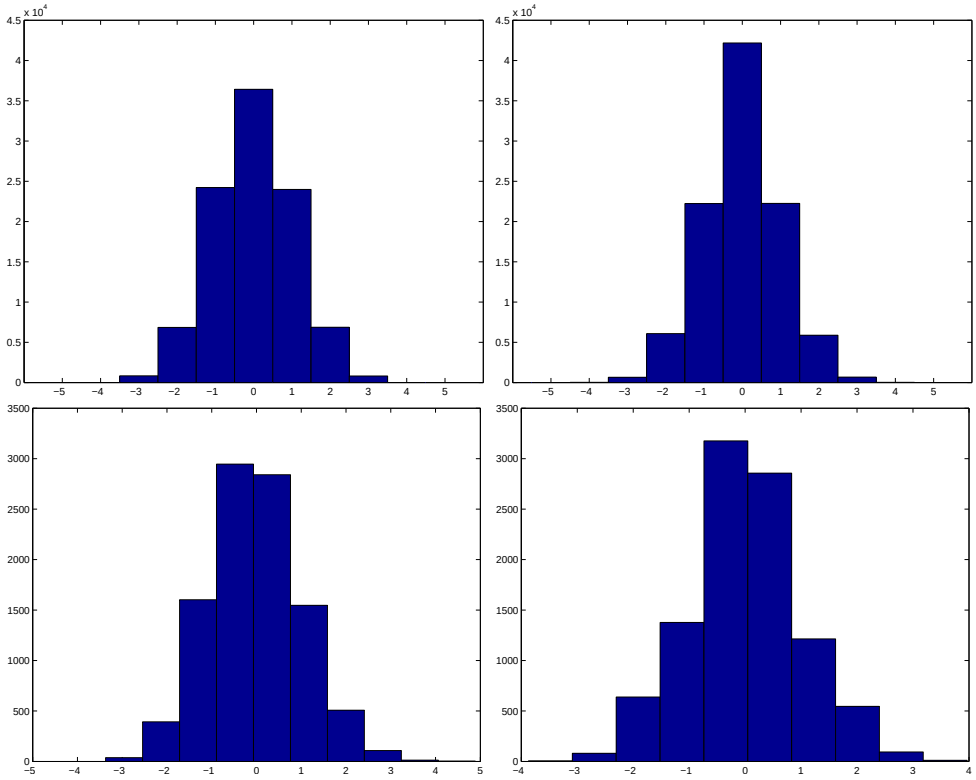

FIGURE 1.20: *The* **top row** *are the normalized histograms from figure 1.18, and the* **center row** *are the normalized histograms from figure 1.19. Notice that these histograms are very difficult to tell apart, even though they describe the results of a variety of experiments.*

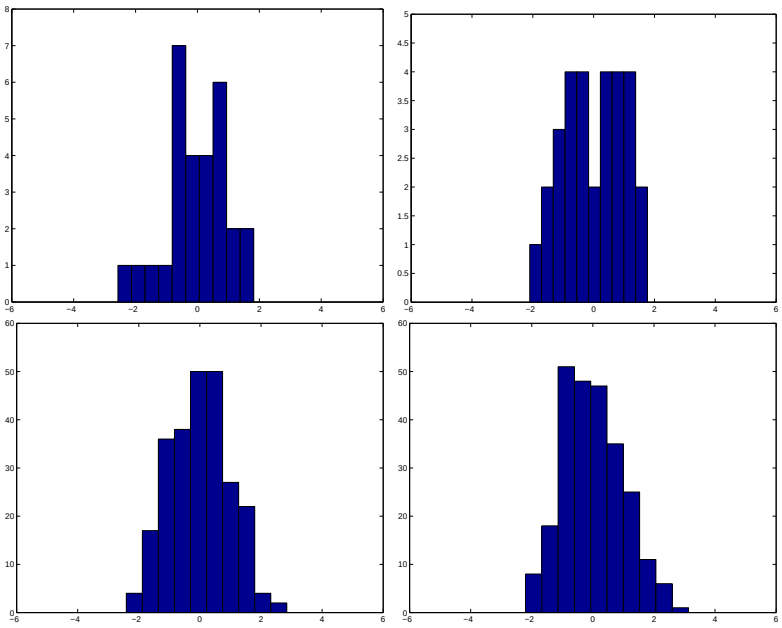

FIGURE 1.21: *Many quite different data sets have histograms that are similar to the normal curve. Each of these histograms has been normalized by subtracting the mean, then dividing by the standard deviation. On the* **top left**, *a histogram of measurements of the density of earth relative to water (from* `http://lib.stat.cmu.edu/DASL/Datafiles/distributiondat.html`*); on the* **top right**, *a histogram of the volumes of 30 oysters (from* `http://www.amstat.org/publications/jse/jse_data_archive.htm`*; look for 30oysters.dat.txt); on the* **bottom left**, *a histogram of human heights (from* `http://www2.stetson.edu/~jrasp/data.htm`*; look for bodyfat.xls, with two outliers removed); and on the* **bottom right**, *a histogram of human weights (from* `http://www2.stetson.edu/~jrasp/data.htm`*; look for bodyfat.xls, with two outliers removed).*

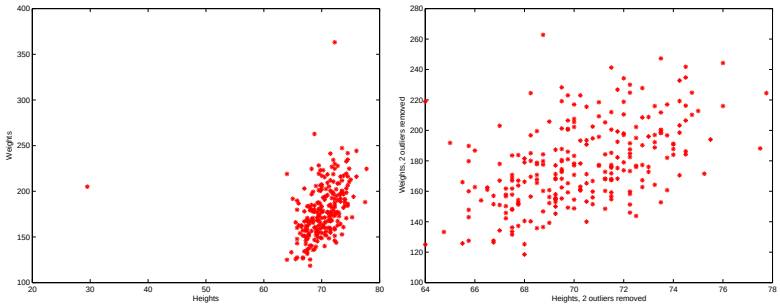

FIGURE 1.22: *On the* **left**, *a scatter plot of height against weight, from the dataset at* `http://www2.stetson.edu/~jrasp/data.htm`; *bodyfat.xls. It is difficult to draw conclusions from this plot, because there are two outliers which dominate the picture. To show the outliers, the rest of the data has had to be bunched up.* **Center**, *a scatter plot of height against weight, with the outliers removed. This suggests, but doesn't conclusively establish, that taller people are heavier and vice versa.*

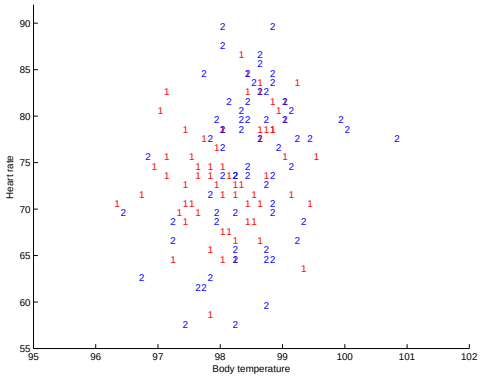

FIGURE 1.23: *On the* **right**, *a scatter plot of body temperature against heart rate, from the dataset at* `http://www2.stetson.edu/~jrasp/data.htm`; *normtemp.xls. I have separated the two genders by plotting a different symbol for each (though I don't know which gender is indicated by which letter); 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 any dependence between temperature and heart rate isn't affected by gender.*

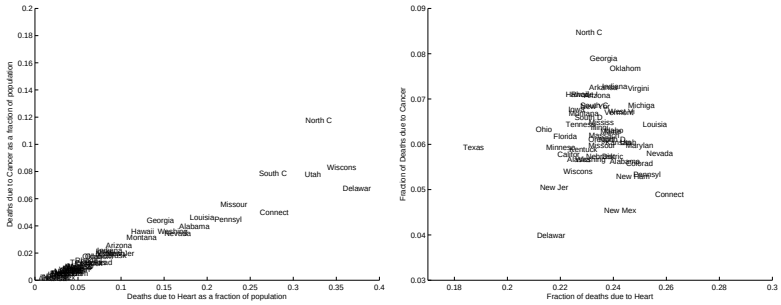

FIGURE 1.24: *On the* **left**, *a scatter plot of deaths from cancer as a fraction of total population against deaths from heart as a fraction of total population, from the deaths dataset at* `http://www2.stetson.edu/~jrasp/data.htm`*; deathdata.xls. I have used the first 7 characters of the state's name as a token here. On the* **right**, *a scatter plot of deaths from cancer as a fraction of all deaths against deaths from heart as a fraction of all deaths, from the deaths dataset at* `http://www2.stetson.edu/~jrasp/data.htm`*; deathdata.xls.*

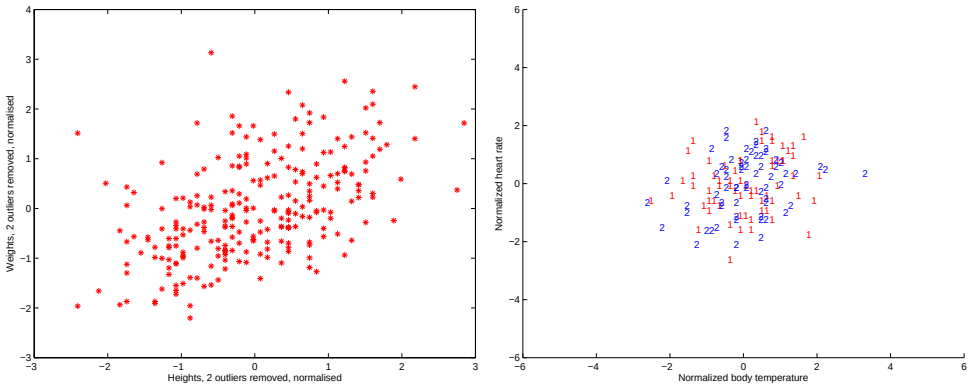

FIGURE 1.25: *Each of these is a normalized scatter plot. The x and y coordinates of each data point have been normalized by subtracting the mean, then dividing by the standard deviation. On the* **left**, *a scatter plot of height and weight from the dataset of* `http://www2.stetson.edu/~jrasp/data.htm`; *bodyfat.xls (with two outliers removed), where each variable has been normalized. Notice how higher individuals tend to have larger weights, because the scatter plot forms an oval leaning to the right. On the* **right**, *a scatter plot of body temperature data against heart rate, from the dataset of* `http://www2.stetson.edu/~jrasp/data.htm`; *normtemp.xls. Again, each variable has been normalized. Notice there is little evidence of a relationship between body temperature and heart rate.*

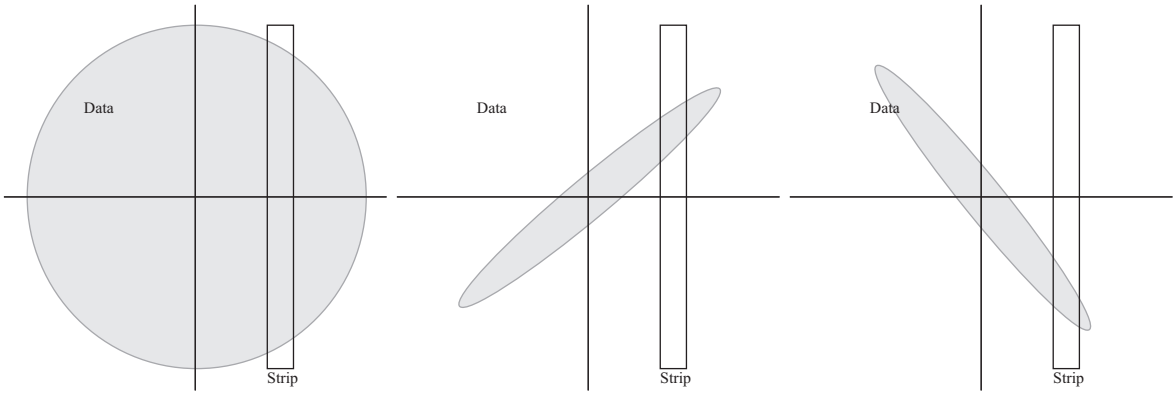

FIGURE 1.26: *On the* **left**, *a sketch of a scatter plot of data in standard coordinates where there is no relationship between $\hat{x}$ and $\hat{y}$. In this case, we expect that, for a small strip of $\hat{x}$, any value of $\hat{y}$ could appear in this strip — all the strips have about the same mean, though the variance changes. This means that knowing $\hat{x}$ for some data point does not help us guess $\hat{y}$. In the text, we show this means the correlation coefficient is about zero. If $\hat{y}$ tends to be large for large $\hat{x}$ and small for small $\hat{x}$, then we see the sketch in the* **center**. *This will result in a positive correlation coefficient. Finally, if $\hat{y}$ tends to be small for large $\hat{x}$ and large for small $\hat{x}$, then we see the sketch on the* **right**. *This will result in a negative correlation coefficient.*

Definition: *Predicting a value using correlation*

Assume we have a dataset of N points (x_i, y_i). 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

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

- Compute the correlation

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

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

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

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

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

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

$$\begin{aligned}\mathsf{mean}\left(\left\{u_i^2\right\}\right) &= \mathsf{mean}\left(\left\{y_i^2\right\}\right) - 2r\mathsf{mean}\,(\{x_i y_i\}) + r^2\mathsf{mean}\left(\left\{x_i^2\right\}\right) \\ &= 1 - 2r^2 + r^2 \\ &= 1 - r^2\end{aligned}$$

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. 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. 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).

1.3.4 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. It DOES NOT mean that changing one variable causes the other to change (sometimes known as causation). It means ONLY that they appear to change together.

Variables could change together for a variety of reasons. One is that they just happen to do so (i.e. you got lucky in your observations — we will quantify that later). Another is that there is some causal relationship — for example, pressing the accelerator tends to make the car go faster, and you'd expect to see some correlation between accelerator depression and car acceleration. Yet another is that there is some other background variable, linked causally to each of the observed variables.

This is best illustrated with examples. 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. 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).

Another nice example comes from Vickers (ibid). The graph, shown in Figure 1.28, 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 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.

1.3.5 Correlation outside Standard Coordinates

It's not always convenient or a good idea to produce scatter plots in standard coordinates (among other things, doing so hides the units of the data, which can be a nuisance). Fortunately, scaling or translating data does not change the value

of the correlation coefficient (though it can change the sign).

One way to see this is that the correlation coefficient 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.

1.3.6 Covariance

Correlation and standard deviation can each be seen as an instance of a more general operation on data. Assume that we have two one dimensional data sets x_i and y_i. Then we can define the **covariance** of x_i and y_i.

Definition: *Covariance*

Assume we have two sets of N data items, x_i and y_i for $i = 1, \ldots, N$, where $N > 1$. Write $\mu_x = \mathsf{mean}\left(\{x_i\}\right)$ and $\mu_y = \mathsf{mean}\left(\{y_i\}\right)$ We compute the covariance by

$$\mathsf{cov}\left(\{x_i\}, \{y_i\}\right) = \frac{\sum_i (\hat{x}_i - \mu_x)(\hat{y}_i - \mu_y)}{N}$$

Covariance is like correlation. It's only really meaningful when the order of the data is meaningful. If you reorder y_i but not x_i, then the covariance you compute will change. The most common case where covariance is meaningful is when x_i and y_i are components of vectors in a dataset. So, for example, x_i could be the x-component and y_i could be the y component of data points on the plane. As another example, x_i could be the 27'th and y_i the 91'st component of a 100 dimensional dataset. We will see more about covariance in this form later.

When the order of the data is meaningful, covariance measures the tendency of x_i and y_i to be larger than (resp. smaller than) the mean at the same time. If x_i tends to be larger (resp. smaller) than its mean for data points where y_i is also larger (resp. smaller) than its mean, then the covariance should be positive. If x_i tends to be larger (resp. smaller) than its mean for data points where y_i is smaller (resp. larger) than its mean, then the covariance should be negative.

Covariance is familiar, in other forms. First, notice that

$$\mathsf{std}\left(x_i\right) = \sqrt{\mathsf{cov}\left(\{x_i\}, \{x_i\}\right)}$$

and second, notice that

$$\mathsf{corr}\left(\{(x_i, y_i)\}\right) = \frac{\mathsf{cov}\left(\{x_i\}, \{y_i\}\right)}{\sqrt{\mathsf{cov}\left(\{x_i\}, \{x_i\}\right)}\sqrt{\mathsf{cov}\left(\{y_i\}, \{y_i\}\right)}}.$$

This is occasionally a useful way to think about correlation. It says that the correlation measures the tendency of x_i and y_i to be larger (resp. smaller) than their means for the same data points, compared to how much they change on their own.

Listing 1.4: Matlab code used to produce boxplots for the pizza data

```
cd('~/Current/Courses/Probcourse/SomeData/DataSets/');
[num, txt, raw]=xlsread('cleanpizzasize.xls');
ndat=size(num, 1);
figure(1); boxplot(num(:, 5), txt(:, 2)); figure(1)
print -depsc2 pmakerboxes.eps
t2=txt(:, 1);
for i=1:ndat
    foo=txt(i, 2);
    bar=foo{1};
    c1=bar(1);
    foo=txt(i, 3);
    bar=foo{1};
    c2=bar(1);
    foo=txt(i, 4);
    bar=foo{1};
    c3=bar(1);
    c4=bar(end);
    t2(i)={strcat(c1, c2, c3, c4)};
end
t3=txt(:, 1);
for i=1:ndat
    foo=txt(i, 2);
    bar=foo{1};
    c1=bar(1);
    t3(i)=strcat(c1, '-', txt(i, 3));
end
figure(1); boxplot(num(:, 5), t3); figure(1);
print -depsc2 ptypeboxes.eps
figure(2); boxplot(num(:, 5), t2, 'grouporder', {'DCBs','DCHn',...
    'DCSe', 'DDBs', 'DDHn', 'DDSe', 'DTBs', 'DTHn', 'DTSe', ...
    'EDBs', 'EDHn', 'EDSo', 'EMBs', 'EMHn', 'EMSo', 'ETBs', ...
    'ETHn', 'ETSo'}); figure(2);
print -depsc2 pdetailboxes.eps
```

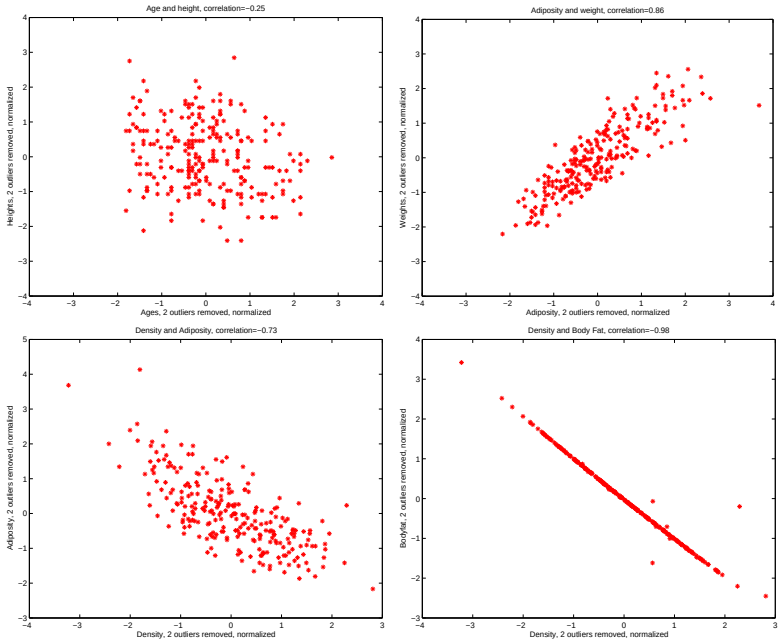

FIGURE 1.27: *Scatter plots for various pairs of variables for the age-height-weight dataset from* `http://www2.stetson.edu/~jrasp/data.htm`*; bodyfat.xls. In each case, two outliers have been removed, and the plots are in standard coordinates. On the* **top left**, *a scatter plot of age against height. In this dataset, these variables are hardly correlated, but younger people tend to be slightly taller. The correlation coefficient is -0.25. On the* **top right**, *a scatter plot of adiposity (not defined in the metadata, though presumably some measure of the amount of fatty tissue) against weight. These variables are fairly naturally correlated, and the correlation coefficient is 0.86. On the* **bottom left**, *a scatter plot of average tissue density against adiposity. 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 on the* **bottom right**, *a scatter plot of density against body weight. The correlation coefficient is -0.98.*

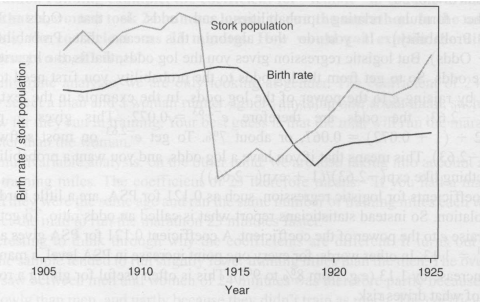

FIGURE 1.28: *This figure, from Vickers (ibid, p184) shows a plot of the stork population as a function of time, and the human birth rate as a function of time, for some years in Germany. The correlation is fairly clear; but this does not mean that reducing the number of storks means there are fewer able to bring babies. Instead, this is the impact of the first world war — a hidden or latent variable.*

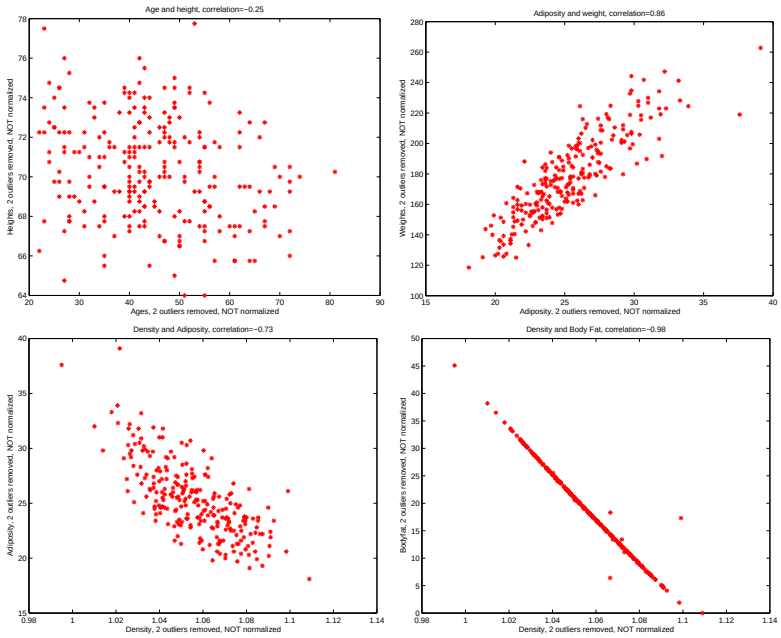

FIGURE 1.29: *Scatter plots for various pairs of variables for the age-height-weight dataset from* `http://www2.stetson.edu/~jrasp/data.htm`*; bodyfat.xls. In each case, two outliers have been removed, but the plots are* not *in standard coordinates. On the* **top left**, *a scatter plot of age against height. In this dataset, these variables are hardly correlated, but younger people tend to be slightly taller. The correlation coefficient is -0.25. On the* **top right**, *a scatter plot of adiposity (not defined in the metadata, though presumably some measure of the amount of fatty tissue) against weight. These variables are fairly naturally correlated, and the correlation coefficient is 0.86. On the* **bottom left**, *a scatter plot of average tissue density against adiposity. 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 on the* **bottom right**, *a scatter plot of density against body weight. The correlation coefficient is -0.98.*