First Tools for Looking at Data

1

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.

1.1 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 1–10—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.

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).

© Springer International Publishing AG 2018

D. Forsyth, *Probability and Statistics for Computer Science*,
https://doi.org/10.1007/978-3-319-64410-3_1

Each element of a tuple has its own type. Some elements might be categorical. For example, one dataset we shall see several times has 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) Gender, or add Age to Grades.

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. 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 two-dimensional data, and we look at high dimensional data in Chap. 10.

1.2 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 1.1 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 1.1 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.

Table 1.2 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 in a paper "The Role of Sports as a Social Determinant for Children," published in *Research Quarterly for Exercise and Sport* in 1992. 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

Table 1.1 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

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

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

Table 1.2 Chase and Dunner, in a study described in the text, collected data on what students thought made other students popular

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

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

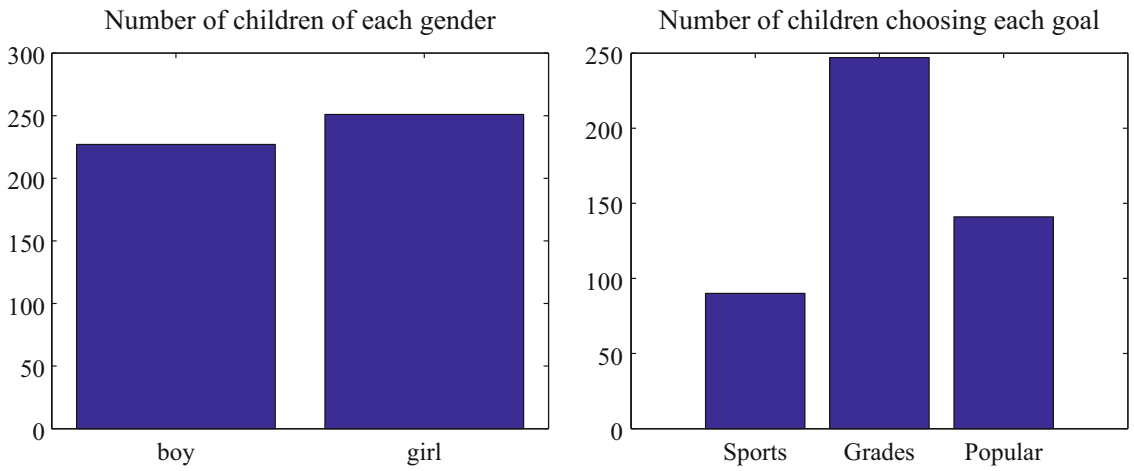


Fig. 1.1 On the *left*, a bar chart of the number of children of each gender in the Chase and Dunner study. Notice that there are about the same number of boys and girls (the bars are about the same height). On the *right*, a bar chart of the number of children selecting each of three goals. You can tell, at a glance, that different goals are more or less popular by looking at the height of the bars

be very hard to read. Table 1.2 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.

1.2.1 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 1.1 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 1.2, because I showed only the first 20 items; but a 478 item table is very difficult to read.

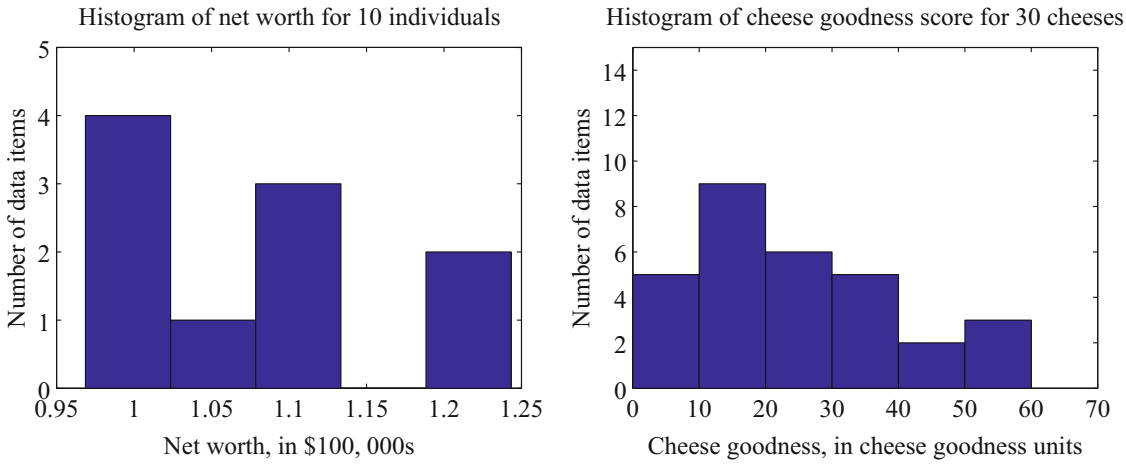


Fig. 1.2 On the *left*, a histogram of net worths from the dataset described in the text and shown in Table 1.1. On the *right*, a histogram of cheese goodness scores from the dataset described in the text and shown in Table 1.1

1.2.2 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.

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 1.2 shows a histogram of the data in Table 1.1. 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 1.2 also shows a histogram of the data in Table 1.1. 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.

1.2.3 How to Make Histograms

Usually, one makes a histogram by finding the appropriate command or routine in your programming environment. I use Matlab and R, depending on what I feel like. It is useful to understand the procedures used to make and plot histograms.

Histograms with Even Intervals: The easiest histogram to build uses equally sized intervals. Write x_i for the i'th number in the dataset, $x_{\min}$ for the smallest value, and $x_{\max}$ for the largest value. We divide the range between the smallest and largest values into n intervals of even width $(x_{\max} - x_{\min})/n$. In this case, the height of each box is given by the number of items in that interval. We could represent the histogram with 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.

Histograms with Uneven Intervals: For a histogram with even intervals, it is natural that 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 Fig. 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. It is natural that the height of the fused box should be the average height of the two boxes. This observation gives us a rule.

Write dx for the width of the intervals; n_1 for the height of the box over the first interval (which is the number of elements in the first box); and n_2 for the height of the box over the second interval. The height of the fused box will be $(n_1 + n_2)/2$. Now the *area* of the first box is $n_1 dx$; of the second box is $n_2 dx$; and of the fused box is $(n_1 + n_2)dx$. For each of these boxes, the *area* of the box is proportional to the number of elements in the box. This gives the correct rule: plot boxes such that the area of the box is proportional to the number of elements in the box.

1.2.4 Conditional Histograms

Most people believe that normal body temperature is 98.4° 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 http://www2.stetson.edu/~jrasp/data.htm. This data appears on Dr. John Rasp's statistics data website, and apparently first came from a paper in the *Journal of Statistics Education*. As you can see from the histogram (Fig. 1.3), 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. The dataset gives genders as 1 or 2—I don't know which is male and which female. Histograms that plot only part of a dataset are sometimes called **conditional histograms** or **class-conditional histograms**, because each histogram is conditioned on something. In this case, each histogram uses only data that comes from a particular gender. Figure 1.3 gives the class conditional histograms. It does seem like individuals of one gender run a little cooler than individuals of the other. Being certain takes considerably more work than looking at these histograms, because the difference might be caused by an unlucky choice of subjects. But the histograms suggests that this work might be worth doing.

1.3 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).

1.3.1 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 1.1 (Mean) Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. Their mean is

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

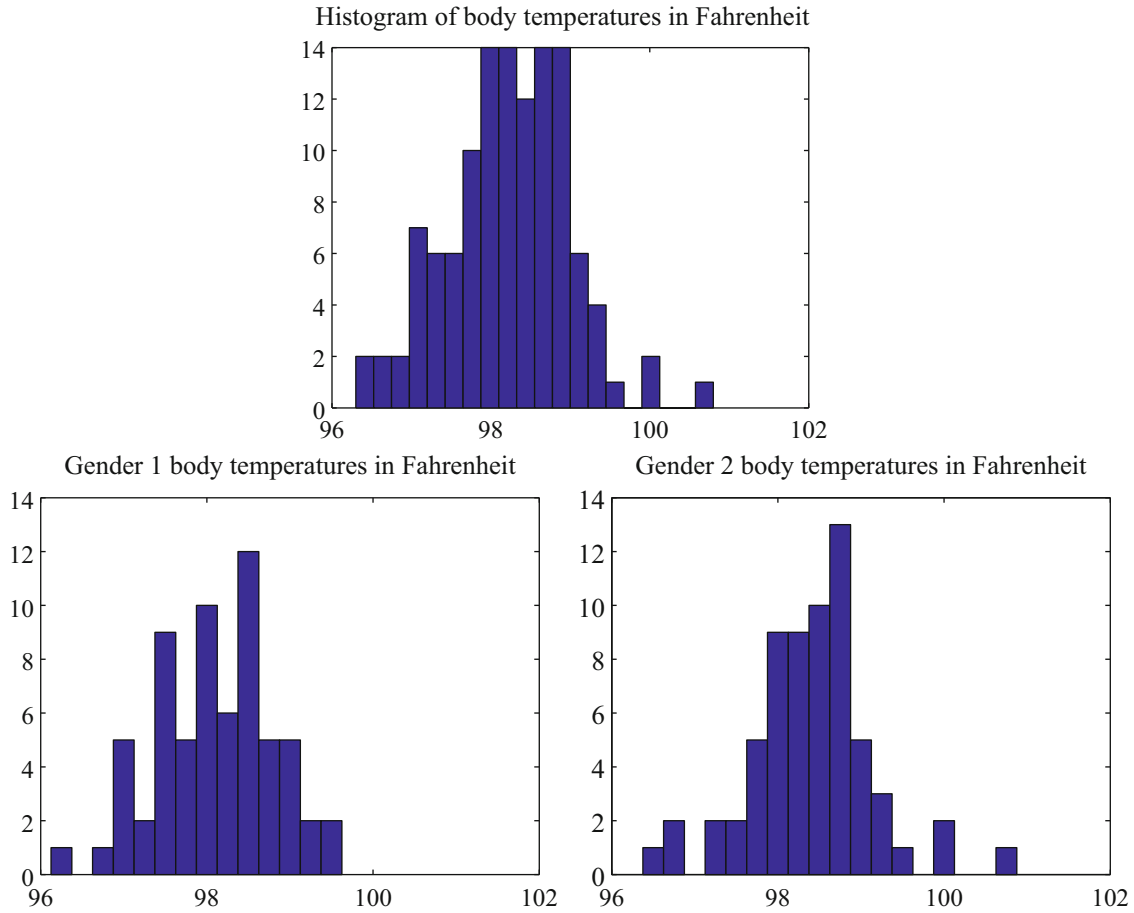


Fig. 1.3 On *top*, a histogram of body temperatures, from the dataset published at http://www2.stetson.edu/~jrasp/data.htm. These seem to be clustered fairly tightly around one value. The *bottom row* shows histograms for each gender (I don't know which is which). It looks as though one gender runs slightly cooler than the other

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.

The mean has several important properties you should remember. These properties are easy to prove (and so easy to remember). I have broken these out into a box of useful facts below, to emphasize them.

Useful Facts 1.1 (Properties of the Mean)

- 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
$$\text{mean}\left(\{x_i + c\}\right) = \text{mean}\left(\{x_i\}\right) + c.$$
- The sum of signed differences from the mean is zero: or,
$$\sum_{i=1}^{N}(x_i - \text{mean}\left(\{x_i\}\right)) = 0.$$

(continued)

- 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

$$\underset{\mu}{\text{argmin}} \sum_i (x_i - \mu)^2 = \text{mean}(\{x_i\}).$$

All this means that the mean is a location parameter; it tells you where the data lies along a number line.

I prove $\underset{\mu}{\text{argmin}} \sum_i (x_i - \mu)^2 = \text{mean}(\{x\})$ below. This result means that the mean is the single number that is closest to all the data items. The mean tells you where the overall blob of data lies. For this reason, it is often referred to as a **location parameter**. If you choose to summarize the dataset with a number that is as close as possible to each data item, the mean is the number to choose. The mean is also a guide to what new values will look like, if you have no other information. For example, in the case of the bar, a new person walks in, and I must guess that person's net worth. Then the mean is the best guess, because it is closest to all the data items we have already seen. In the case of the bar, if a new person walked into this bar, and you had to guess that person's net worth, you should choose $107,903.

Property 1.1 The Average Squared Distance to the Mean is Minimized

Proposition $\underset{\mu}{\textit{argmin}} \sum_i (x_i - \mu)^2 = \text{mean}(\{x\})$

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:

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

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

1.3.2 Standard Deviation

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

(continued)

$$\begin{aligned}\text{std}(\{x_i\}) &= \sqrt{\frac{1}{N}\sum_{i=1}^{i=N}(x_i - \text{mean}(\{x\}))^2} \\ &= \sqrt{\text{mean}(\{(x_i - \text{mean}(\{x\}))^2\})}.\end{aligned}$$

You should think of the standard deviation as a scale. It measures the size of the average deviation from the mean for a dataset, or how wide the spread of data is. For this reason, it is often referred to as a **scale parameter**. 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 deviations 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

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

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

$$\text{abs}(x_i - \text{mean}(\{x\})) > k\text{std}(\{x\}).$$

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. Standard deviation has very important properties. Again, for emphasis, I have broken these properties out in a box below.

Useful Facts 1.2 (Properties of Standard Deviation)

- Translating data does not change the standard deviation, i.e. $\text{std}(\{x_i + c\}) = \text{std}(\{x_i\})$.
- Scaling data scales the standard deviation, i.e. $\text{std}(\{kx_i\}) = k\text{std}(\{x_i\})$.
- For any dataset, there can be only a few items that are many standard deviations away from the mean. For N data items, x_i, whose standard deviation is σ, 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, that is, $(\text{std}(\{x\}))^2 \leq \max_i(x_i - \text{mean}(\{x\}))^2$.

The standard deviation is often referred to as a scale parameter; it tells you how broadly the data spreads about the mean.

Property 1.2 For any dataset, it is hard for data items to get many standard deviations away from the mean.

Proposition *Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. Assume the standard deviation of this dataset is $\text{std}(\{x\}) = \sigma$. 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. Now we must 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

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

(continued)

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

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

for any kind of data at all.

The bound in proof 1.2 is true for *any kind of data*. 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. 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% is within three standard deviations of the mean, and the percentage of data that is within (say) ten standard deviations of the mean is essentially indistinguishable from 100%.

Property 1.3 For any dataset, there must be at least one data item that is at least one standard deviation away from the mean.

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

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

The properties proved in proof 1.2 and proof 1.3 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.

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

$$\text{stdunbiased}\left(\{x\}\right) = \sqrt{\frac{\sum_i (x_i - \text{mean}\left(\{x\}\right))^2}{N-1}}$$

(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 whether I know what I'm talking about. In this case, I do (although I would say that).

The confusion arises because sometimes the datasets we see are actually samples of larger datasets. For example, in some circumstances you could think of the net worth dataset as a sample of all the net worths in the USA. In such cases, we are often interested in the standard deviation *of the underlying dataset that was sampled* (rather than of the dataset of samples that you have). The second number is a slightly better way to estimate this standard deviation than the definition we have been working with. Don't worry—the N in our expressions is the right thing to use for what we're doing.

1.3.3 Computing Mean and Standard Deviation Online

One useful feature of means and standard deviations is that you can estimate them online. Assume that, rather than seeing N elements of a dataset in one go, you get to see each one once in some order, and you cannot store them. This means that after seeing k elements, you will have an estimate of the mean based on those k elements. Write $\hat{\mu}_k$ for this estimate. Because

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

and

$$\sum_{i=1}^{k+1} x_i = \left(\sum_{i=1}^{k} x_i\right) + x_{k+1},$$

we have the following recursion

$$\hat{\mu}_{k+1} = \frac{(k\hat{\mu}_k) + x_{k+1}}{(k+1)}.$$

Similarly, after seeing k elements, you will have an estimate of the standard deviation based on those k elements. Write $\hat{\sigma}_k$ for this estimate. We have the recursion

$$\hat{\sigma}_{k+1} = \sqrt{\frac{(k\hat{\sigma}_k^2) + (x_{k+1} - \hat{\mu}_{k+1})^2}{(k+1)}}.$$

1.3.4 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 1.3 (Variance) Assume we have a dataset $\{x\}$ of N data items, $x_1, \ldots, x_N$. where $N > 1$. Their variance is:

$$\begin{aligned}\text{var}(\{x\}) &= \frac{1}{N}\left(\sum_{i=1}^{i=N}(x_i - \text{mean}(\{x\}))^2\right) \\ &= \text{mean}\left(\left\{(x_i - \text{mean}(\{x\}))^2\right\}\right).\end{aligned}$$

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. The properties of the variance follow from the fact that it is the square of the standard deviation. I have broken these out in a box, for emphasis.

Useful Facts 1.3 (Properties of Variance)

- $\text{var}(\{x + c\}) = \text{var}(\{x\})$.
- $\text{var}(\{kx\}) = k^2\text{var}(\{x\})$.

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!).

1.3.5 The Median

One problem with the mean is that it can be affected strongly by extreme values. Go back to the bar example, of Sect. 1.3.1. 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 1.4 (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

$$\text{median}(\{x\})$$

for the operator that returns the median.

For example,

$$\begin{aligned}\text{median}(\{3,5,7\}) &= 5, \\ \text{median}(\{3,4,5,6,7\}) &= 5,\end{aligned}$$

and

$$\text{median}\,(\{3,4,5,6\}) = 4.5.$$

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,

$$\text{median}\,(\{1,2,2,2,2,2,2,2,3\}) = 2.$$

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. You can think of the median of a dataset as giving the "middle" or "center" value. It is another way of estimating where the dataset lies on a number line (and so is another location parameter). 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, which I have broken out in a box. Each is easily proved, and proofs are relegated to the exercises.

Useful Facts 1.4 (Properties of the Median)

- $\text{median}\,(\{x+c\}) = \text{median}\,(\{x\}) + c$.
- $\text{median}\,(\{kx\}) = k\text{median}\,(\{x\})$.

1.3.6 Interquartile Range

Outliers are a nuisance in all sorts of ways. Plotting the histogram of the net worth data with the billionaire included will be tricky. Either you leave the billionaire out of the plot, or all the histogram bars are tiny. Visualizing this plot shows 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, the mean is about $91M\$$ and the standard deviation is about $300M\$$—so all but one of the data items lie about a third of a standard deviation away from the mean on the small side. The other data item (the billionaire) is about three standard deviations away from the mean on the large side. 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 as a description of the data.

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 $\$3.014 \times 10^8$ isn't really 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 1.5 (Percentile) The k'th percentile is the value such that $k\%$ of the data is less than or equal to that value. We write $\text{percentile}(\{x\}, k)$ for the k'th percentile of dataset $\{x\}$.

Definition 1.6 (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. $\text{percentile}(\{x\}, 25)$). 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. $\text{percentile}(\{x\}, 50)$). 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. $\text{percentile}(\{x\}, 75)$).

Definition 1.7 (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 \$12,350; with the billionaire, it is \$17,710. 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, which I have broken out into a box. Each is easily proved, and proofs are relegated to the exercises.

Useful Facts 1.5 (Properties of the Interquartile Range)

- $\mathsf{iqr}\{x + c\} = \mathsf{iqr}\{x\}$.
- $\mathsf{iqr}\{kx\} = k\mathsf{iqr}\{x\}$.

For most datasets, interquartile ranges tend to be somewhat larger than standard deviations. This isn't really a problem. Each is a method for estimating the scale of the data—the range of values above and below the mean that you are likely to see. It is neither here nor there if one method yields slightly larger estimates than another, as long as you don't compare estimates across methods.

1.3.7 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 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 (in the same book) 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 min. 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 may suffer for it.

1.4 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 (Sect. 1.4.1). Quite remarkably, many different datasets have histograms that have about the same shape (Sect. 1.4.2). 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 1.4.3 describes a clever plot of various summaries of datasets that makes it easier to compare many cases.

1.4.1 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 the case where there are two peaks (Fig. 1.4). 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 (Fig. 1.5). 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 (Fig. 1.6). These are very different from

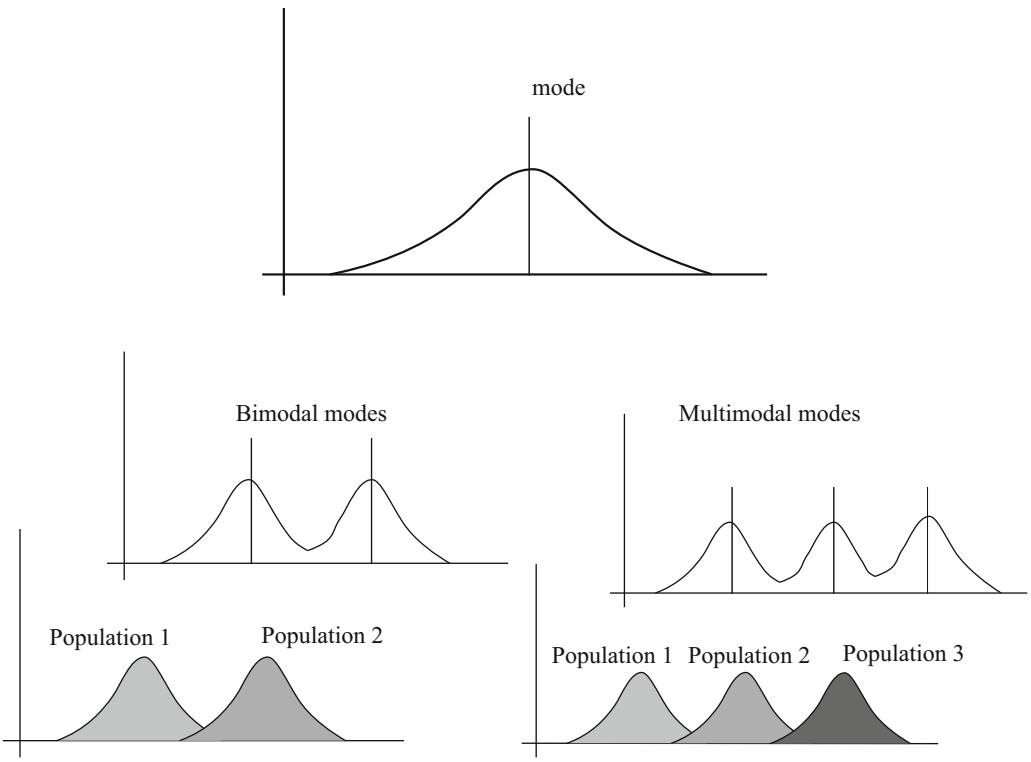


Fig. 1.4 Many histograms are unimodal, like the example on the *top*; there is one peak, or mode. Some are bimodal (two peaks; *bottom left*) or even multimodal (two or more peaks; *bottom right*). One common reason (but not the only reason) is that there are actually two populations being conflated 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.)

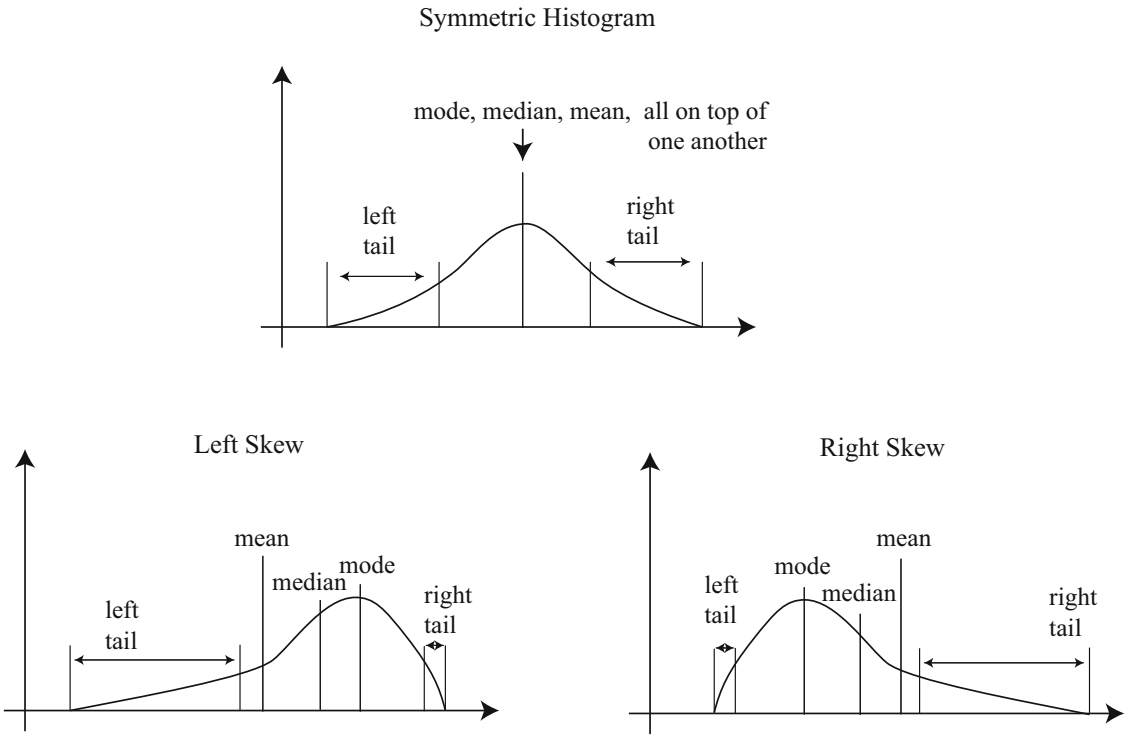


Fig. 1.5 On the *top*, an example of a symmetric histogram, showing its tails (relatively uncommon values that are significantly larger or smaller than the peak or mode). *Lower left*, a sketch of a left-skewed histogram. Here there are few large values, but some very small values that occur with significant frequency. We say the left tail is "long", and that the histogram is left skewed. You may find this confusing, because the main bump is to the right—one way to remember this is that the left tail has been stretched. *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

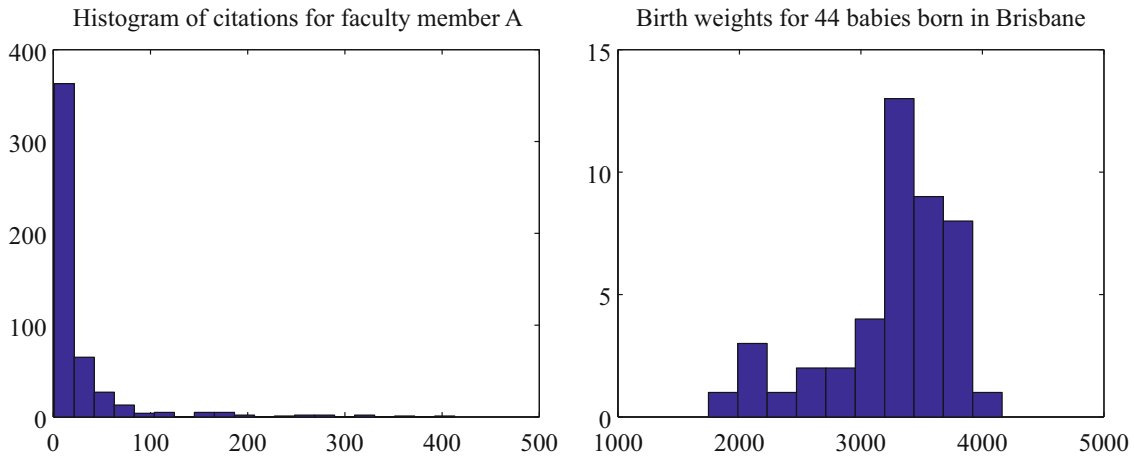


Fig. 1.6 On the *left*, a histogram of citations for a faculty member, from data at http://wiki.stat.ucla.edu/socr/index.php/SOCR_Data_Dinov_072108_H_Index_Pubs. Very few publications have many citations, and many publications have few. This means the histogram is strongly right-skewed. On the *right*, a histogram of birth weights for 44 babies borne in Brisbane in 1997. This histogram looks slightly left-skewed

(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 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; Fig. 1.6 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 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.

1.4.2 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 1.8 (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

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

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

$$\text{mean}(\{\hat{x}\}) = 0.$$

We also have that

$$\text{std}(\{\hat{x}\}) = 1.$$

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 and unimodal, and it looks like a bump. If there were enough data points and the histogram boxes were small enough, the histogram would look like the curve in Fig. 1.7. This phenomenon is so important that data of this form has a special name.

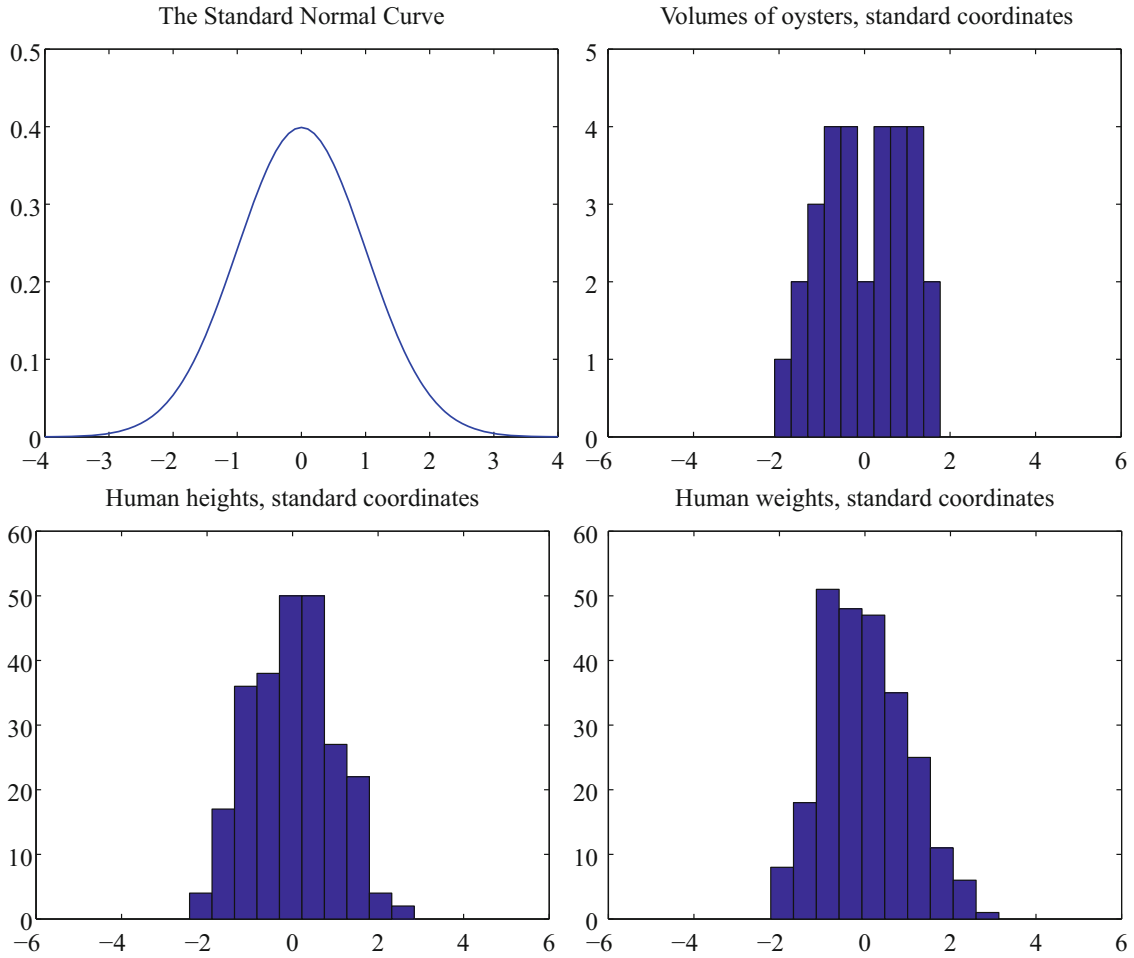


Fig. 1.7 Data is standard normal data when its histogram takes a stylized, bell-shaped form, plotted above. One usually requires a lot of data and very small histogram boxes for this form to be reproduced closely. Nonetheless, the histogram for normal data is unimodal (has a single bump) and is symmetric; 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 1.9 (Standard Normal Data) Data is **standard normal data** if, when we have a great deal of data, the histogram of the data in standard coordinates is a close approximation to the **standard normal curve**. This curve is given by

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

(which is shown in Fig. 1.7).

Definition 1.10 (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 1.7 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, and notice that I removed two outliers); and human weights (from http://www2.stetson.edu/~jrasp/data.htm; look for bodyfat.xls, again, I removed two outliers).

For the moment, assume we know that a dataset is normal. Then we expect it to have the properties in the following box. 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.

Useful Facts 1.6 (Properties of Normal Data)

- 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.

1.4.3 Box Plots

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.

A **box plot** is a way to plot data that simplifies comparison. A box plot 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* (Fig. 1.8).

To build a box plot, 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 complicated, any reasonable programming environment will have a function that will do it for you. Figure 1.8 shows an example box plot. Notice that the rich graphical structure means it is quite straightforward to compare two histograms.

1.5 Whose is Bigger? Investigating Australian Pizzas

At http://www.amstat.org/publications/jse/jse_data_archive.htm), there is 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 Fig. 1.9. 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,

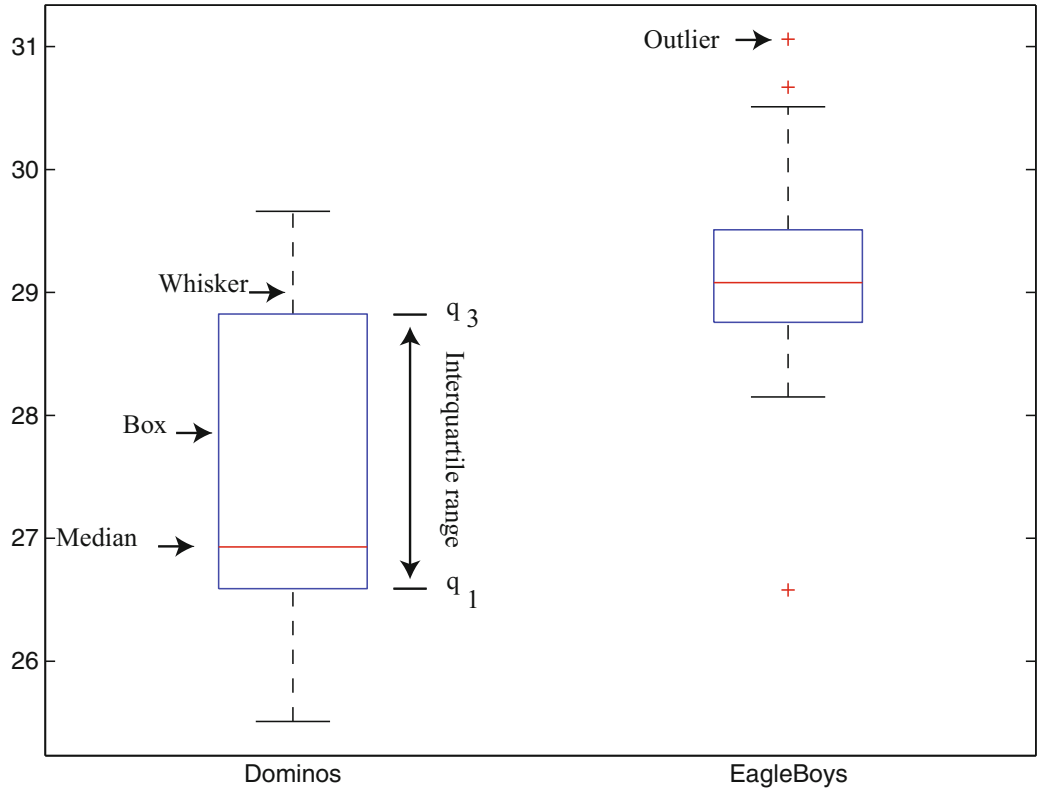
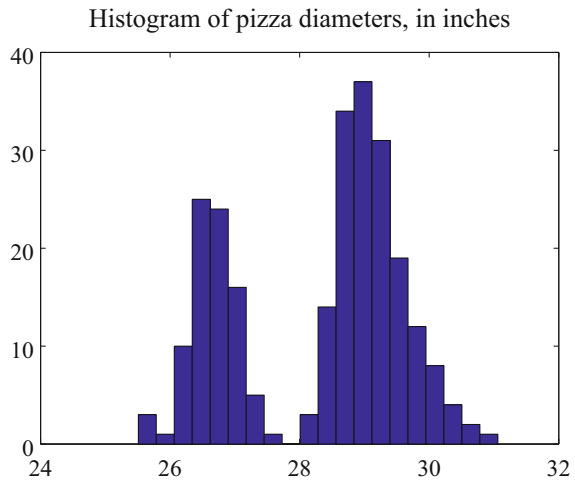


Fig. 1.8 A box plot showing the box, the median, the whiskers and two outliers. Notice that we can compare the two datasets rather easily; the next section explains the comparison

Fig. 1.9 A histogram of pizza diameters from the dataset described in the text. Notice that there seem to be two populations



rather narrow, bump about a mean. This is not what we see in Fig. 1.9—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 Fig. 1.10. 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 box plot (Fig. 1.11), which shows the range

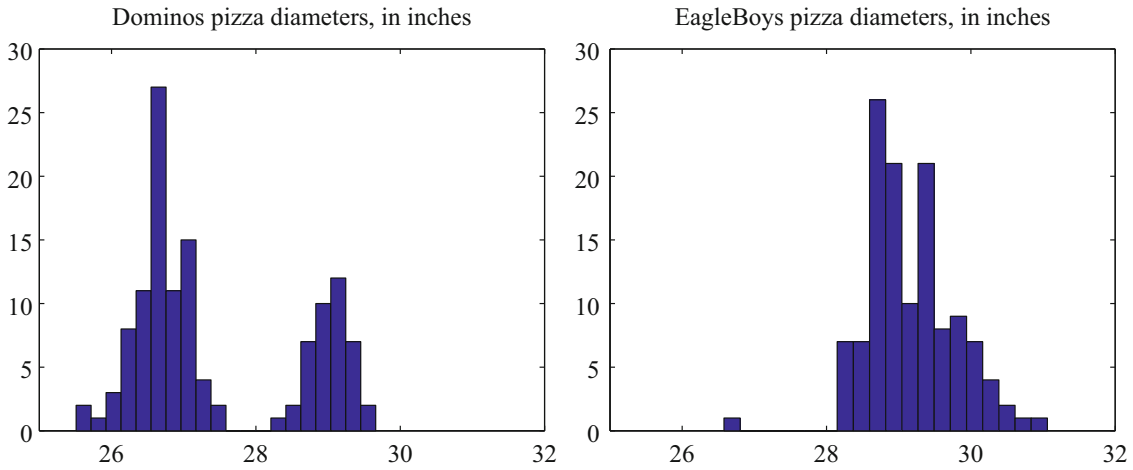


Fig. 1.10 On the *left*, the class-conditional histogram of Dominos pizza diameters from the pizza data set; on the *right*, the class-conditional histogram of EagleBoys pizza diameters. Notice that EagleBoys pizzas seem to follow the pattern we expect—the diameters are clustered tightly around a mean, 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

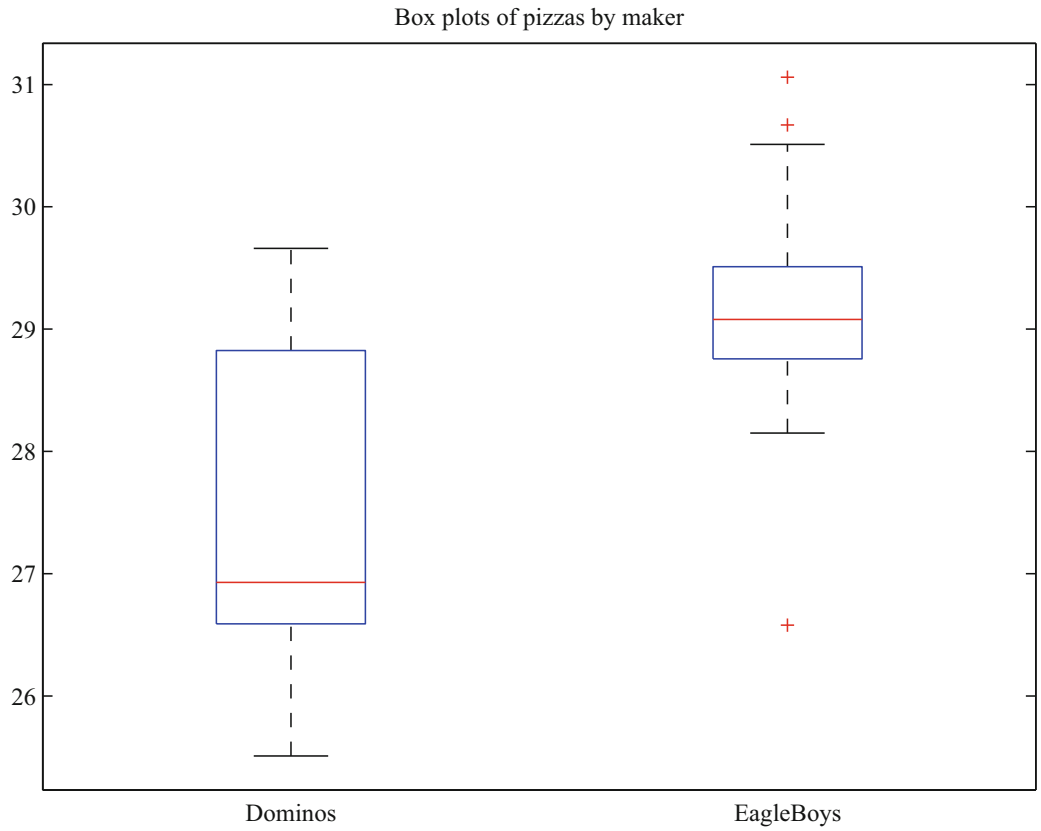


Fig. 1.11 Box Plots of the pizza data, comparing EagleBoys and Dominos pizza. There are several curiosities here: why is the range for Dominos so large (25.5–29)? EagleBoys has a smaller range, but has several substantial outliers; why? One would expect pizza manufacturers to try and control diameter fairly closely, because pizzas that are too small present risks (annoying customers; publicity; hostile advertising) and pizzas that are too large should affect profits

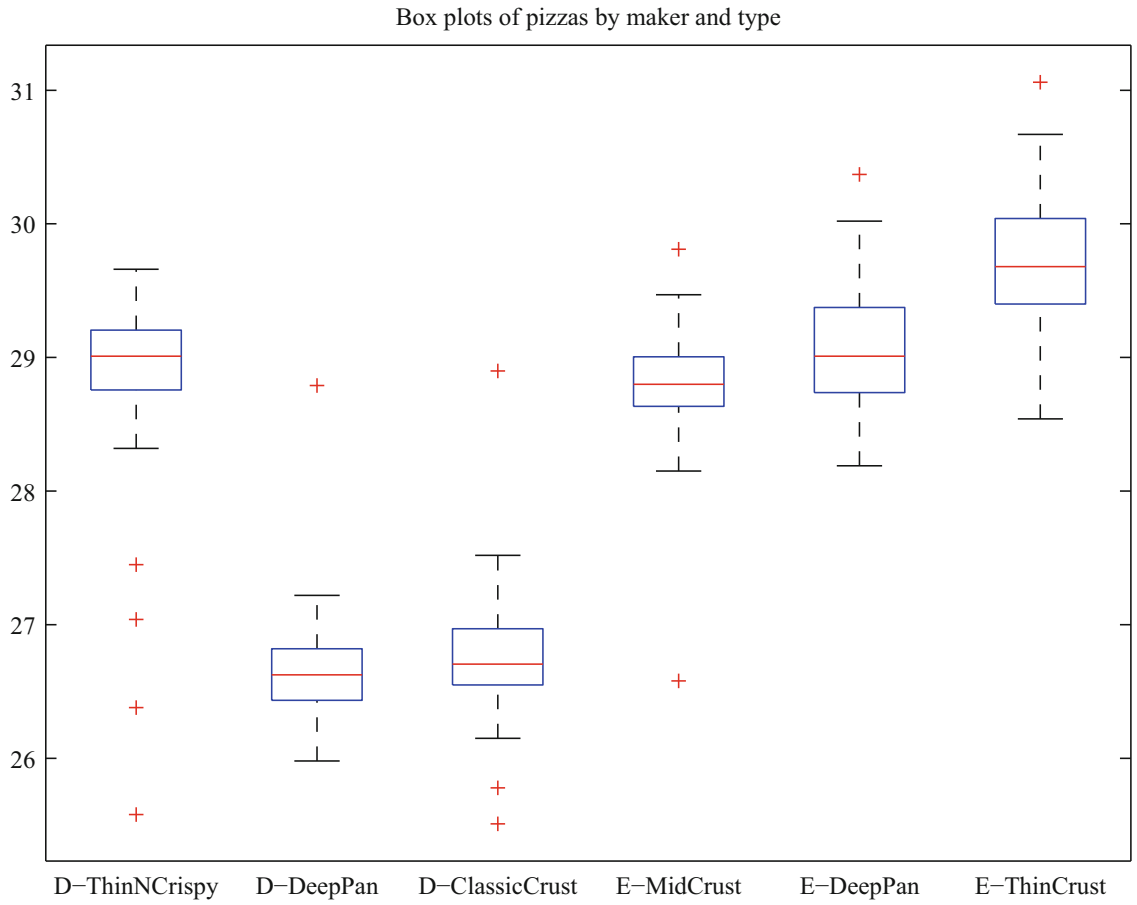


Fig. 1.12 Box Plots for the pizza data, broken out by type (thin crust, etc.)

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 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 box plot is the right way to compare these cases (Fig. 1.12). The box plot 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 box plots (i.e. the diameters for each type and each topping). This leads to rather a lot of boxes (Fig. 1.13), 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 1.13 suggests that the topping isn't what is important, but the crust (group the box plots by eye).

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

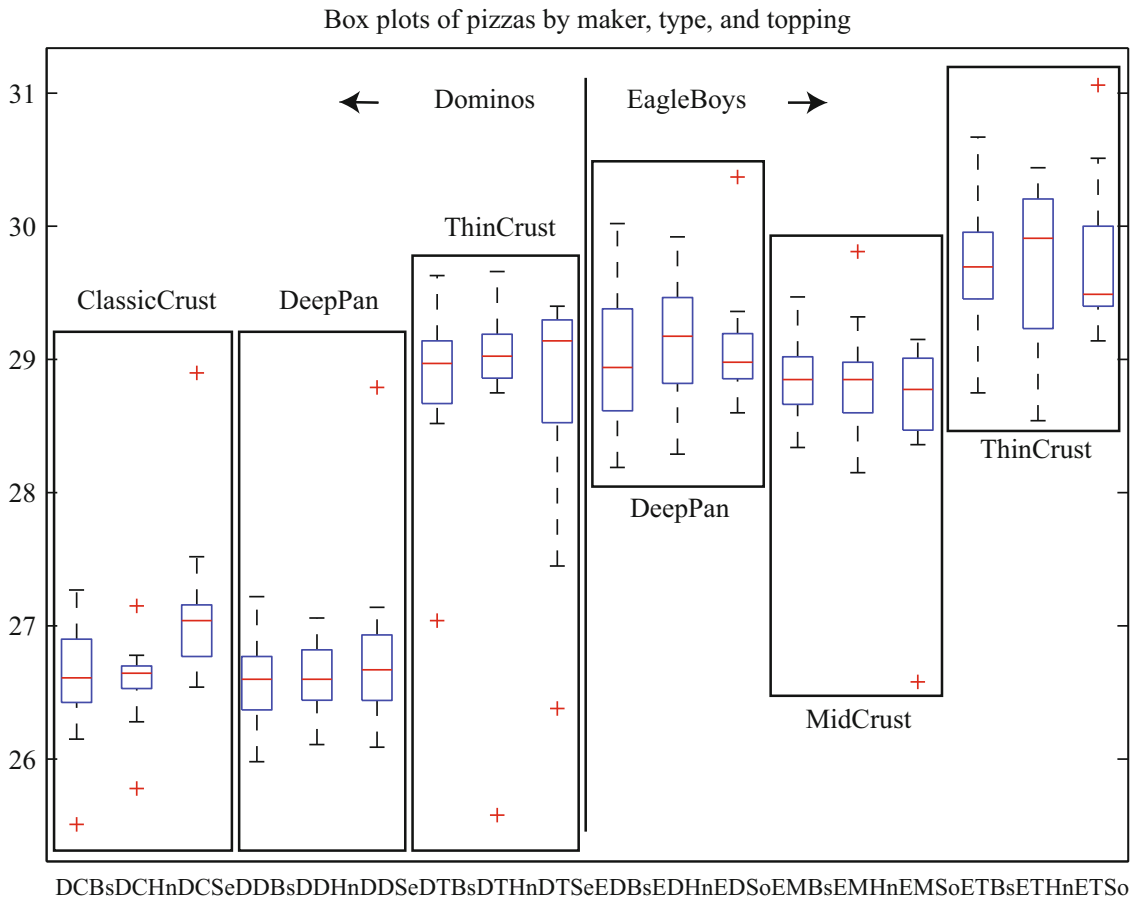


Fig. 1.13 The pizzas are now broken up by topping as well as crust type (look at the source for the meaning of the names). I have separated Dominos from Eagleboys with a vertical line, and grouped each crust type with a box. It looks as though the issue is not the type of topping, but the crust. Eagleboys seems to have tighter control over the size of the final pizza

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 next time). 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.

1.6 You Should

1.6.1 Remember These Definitions

Mean 7
Standard deviation 9
Variance 13
Median 13

Percentile . . . 14
Quartiles . . . 14
Interquartile Range . . . 14
Standard coordinates . . . 18
Standard normal data . . . 19
Normal data . . . 19

1.6.2 Remember These Terms

categorical . . . 3
ordinal . . . 3
continuous . . . 3
bar chart . . . 5
histogram . . . 6
conditional histograms . . . 7
class-conditional histograms . . . 7
average . . . 7
location parameter . . . 9
scale parameter . . . 10
outlier . . . 13
tails . . . 16
mode . . . 16
unimodal . . . 16
multimodal . . . 16
bimodal . . . 16
skew . . . 16
standard normal curve . . . 19
box plot . . . 20

1.6.3 Remember These Facts

Properties of the mean . . . 8
Properties of standard deviation . . . 10
Properties of variance . . . 13
Properties of the median . . . 14
Properties of the interquartile range . . . 15
Properties of normal data . . . 20

1.6.4 Be Able to

- Plot a bar chart for a dataset.
- Plot a histogram for a dataset.
- Tell whether the histogram is skewed or not, and in which direction.
- Plot and interpret conditional histograms.
- Compute basic summaries for a dataset, including mean, median, standard deviation and interquartile range.
- Plot a box plot for one or several datasets.
- Interpret a box plot.
- Use histograms, summaries and box plots to investigate datasets.

Problems

1.1 Show that $\mathsf{mean}\,(\{kx\}) = k\mathsf{mean}\,(\{x\})$ by substituting into the definition.

1.2 Show that $\mathsf{mean}\,(\{x + c\}) = \mathsf{mean}\,(\{x\}) + c$ by substituting into the definition.

1.3 Show that $\sum_{i=1}^{N}(x_i - \mathsf{mean}\,(\{x\})) = 0$ by substituting into the definition.

1.4 Show that $\mathsf{std}\,(\{x + c\}) = \mathsf{std}\,(\{x\})$ by substituting into the definition (you'll need to recall the properties of the mean to do this).

1.5 Show that $\mathsf{std}\,(\{kx\}) = k\mathsf{std}\,(\{x\})$ by substituting into the definition (you'll need to recall the properties of the mean to do this).

1.6 Show that $\mathsf{median}\,(\{x+c\})=\mathsf{median}\,(\{x\}) + c$ by substituting into the definition.

1.7 Show that $\mathsf{median}\,(\{kx\}) = k\mathsf{median}\,(\{x\})$ by substituting into the definition.

1.8 Show that $\mathsf{iqr}\{x + c\} = \mathsf{iqr}\{x\}$ by substituting into the definition.

1.9 Show that $\mathsf{iqr}\{kx\} = k\mathsf{iqr}\{x\}$ by substituting into the definition.

Programming Exercises

1.10 You can find a data set showing the number of barrels of oil produced, per year for the years 1880–1984 at http://lib.stat.cmu.edu/DASL/Datafiles/Oilproduction.html. Is a mean a useful summary of this dataset? Why?

1.11 You can find a dataset giving the cost (in 1976 US dollars), number of megawatts, and year of construction of a set of nuclear power plants at http://lib.stat.cmu.edu/DASL/Datafiles/NuclearPlants.html.

(a) Are there outliers in this data?
(b) What is the mean cost of a power plant? What is the standard deviation?
(c) What is the mean cost per megawatt? What is the standard deviation?
(d) Plot a histogram of the cost per megawatt. Is it skewed? Why?

1.12 You can find a dataset giving the sodium content and calorie content of three types of hot dog at http://lib.stat.cmu.edu/DASL/Datafiles/Hotdogs.html. The types are Beef, Poultry, and Meat (a rather disturbingly vague label). Use class-conditional histograms to compare these three types of hot dog with respect to sodium content and calories.

1.13 You will find a dataset giving (among other things) the number of 3 or more syllable words in advertising copy appearing in magazines at http://lib.stat.cmu.edu/DASL/Datafiles/magadsdat.html. The magazines are grouped by the education level of their readers; the groups are 1, 2, and 3 (the variable is called GRP in the data).

(a) Use a box plot to compare the number of three or more syllable words for the ads in magazines in these three groups. What do you see?
(b) Use a box plot to compare the number of sentences appearing in the ads in magazines in these three groups. What do you see?

1.14 You can find a dataset recording a variety of properties of secondary school students in Portugal athttp://archive.ics.uci.edu/ml/datasets/STUDENT+ALCOHOL+CONSUMPTION. This dataset was collected by P. Cortez and A. Silva, and is hosted by the UC Irvine Machine Learning Repository. There are two datasets; one for students in a math course, and another for students in a Portugese language course.

(a) Use plots of conditional histograms to investigate whether math students drink more alcohol during the week than Portugese language students.
(b) Use plots of conditional histograms to investigate whether students from small families drink more alcohol at the weekend than those from large families.
(c) Each of the variables school, sex, famsize and romantic has two possible values. This means that if we characterize students by the values of these variables, there are sixteen possible types of student. Use box plots to investigate which of these types drinks more alcohol in total.

1.15 You can find a dataset recording some properties of Taiwanese credit card holders at http://archive.ics.uci.edu/ml/datasets/default+of+credit+card+clients. This dataset was collected by I-Cheng Yeh, and is hosted by the UC Irvine Machine Learning Repository. There is a variable indicating whether a holder defaulted or not, and a variety of other variables.

(a) Use plots of conditional histograms to investigate whether people who default have more debt (use the variable X1 for debt) than those who don't default.
(b) Use box plots to investigate whether gender, education or marital status has any effect on the amount of debt (again, use X1 for debt).

1.16 You will find a dataset giving the effects of three poisons and four antidotes at http://www.statsci.org/data/general/poison.html. This dataset records survival times of animals poisoned with one of three poisons and supplied with one of four antidotes. There is no detail supplied on species, protocol, ethical questions, etc.

(a) Use box plots to investigate whether the poisons have different effects for antidote 1.
(b) Use box plots to investigate whether the antidote has any effect for poison 2.