Contents

CHAPTER 1

Inferring information about populations from samples

Inference is the process of looking at the outcomes of experiments, and then determining some underlying facts using some form of model. This is an example of **inference** — we observe some random phenomena, and must then draw conclusions.

Example: *Patriot missiles*

I got this example from "Dueling idiots", a nice book by P.J. Nahin, Princeton University Press. Apparently in 1992, the Boston Globe of Jan 24 reported on this controversy. The pentagon claimed that the patriot missile successfully engaged SCUD missiles in 80% of encounters. An MIT physicist, Theodore Postol, pointed out there was a problem. He viewed tapes of 14 patriot/SCUD encounters, with one hit and 13 misses. We can reasonably assume each encounter is independent. The probability of getting one hit and 13 misses if $P(\text{hit}) = 0.8$ is

$$\left(\begin{array}{c} 14 \\ 1 \end{array} \right) 0.2^{13} 0.8^{1}$$

which is around 1e-8. Now you could look at this data and make several arguments: (a) the probability is 0.8, and the pentagon just got unlucky with the videotapes that Postol looked at; (b) the probability is not 0.8, because you would need to fire 14 patriots at 14 SCUD missiles about 1e8 times to see this set of videotapes once; (c) for some reason, the videotapes are not independent — perhaps only unsuccessful encounters get filmed.

Example: *MTG and Shuffling*

You build a deck of 24 lands and 36 spells. You shuffle this deck, and draw a hand of seven cards. You get no lands. You repeat the experiment, and still get no lands. On a third try, you still get no lands. By the results in chapter 2, this event (three shuffles and draws with no lands in each hand) has probability about 8e-6. You could conclude that the shuffling is not randomizing the cards effectively; the cards might stick together, or you might be bad at shuffling.

Very often the data we see is a small part of the data we could have seen, if we'd been able to collect enough data. We need to know how the measurements we make on the dataset relate to the measurements we could have made, if we had all the data. This situation occurs very often. For example, imagine we wish to know, on a scale of 1-5, how much people like using a touch interface. Asking everyone on the planet and then averaging the answers would be absurd. Instead, we ask a small set of people, chosen rather carefully. If we have chosen sufficiently carefully, then the answer from the small set is quite a good representation of the answer from the whole set.

This gives us a powerful and quite general way of thinking about data. The data we could have observed, if we could have seen everything, is the **population**. The data we actually have is the **sample**. We would like to know the mean of the population, but can see only the sample; surprisingly, we can say a great deal from the sample alone, assuming that it is chosen appropriately.

This framework allows us to use samples to summarize populations; to test a sample to tell whether a population has a particular property; and to ask if two samples represent the same, or different, populations.

1.1 SAMPLES AND POPULATIONS

Assume we have a population $\{x_i\}$, for $i = 1, \ldots, N_p$. Notice the subscript here — this is the number of items in the population, for example, all the people in the world. We want to know the mean of this dataset, but we do not get to see the whole dataset. Instead, we see the sample. This is obtained by choosing a fixed number k, which we expect is a lot smaller than N_p, of data items. Each choice is independent, and fair, meaning that each time we choose, we choose one from the entire set of N_p data items, and each has the same probability of being chosen. This is sometimes referred to as “sampling with replacement”. One model that people often use is to imagine the data items as being written on tickets, which are placed in a jar. You repeat the following experiment k times: shake the jar; take a ticket from the jar and write down the data on the ticket; put it back in the jar. Sometimes the jar is referred to as an “urn”.

1.1.1 Describing the population from a sample

We summarize the whole dataset with a mean, which we write $\mathsf{popmean}\left(\{x_i\}\right)$. The notation is just to drive home the facts that it’s the mean of the whole population, and that we don’t, and can’t, know it. The whole point of this exercise is to estimate this mean.

We would like to estimate the mean of the whole dataset from the items that we actually see. Think about the random variable X whose value is obtained by drawing a ticket from the jar. Write $\mathbb{E}[X]$ for the expected value of a ticket drawn from the jar. Then we have that

$$\begin{aligned}
\mathbb{E}[X] &= \sum_{i\in 1,\ldots N_p} x_i p(i) \\
&= \sum_{i\in 1,\ldots N_p} x_i \frac{1}{N_p} && \text{because we draw fairly from the jar} \\
&= \frac{\sum_{i\in 1,\ldots N_p} x_i}{N_p} \\
&= \mathsf{popmean}\left(\{x_i\}\right)
\end{aligned}$$

which is the mean value of the items in the jar. Now imagine we draw k tickets from the jar, and average them. Write $X^{(k)}$ for this random variable. We must have that

$$\mathbb{E}\left[X^{(k)}\right] = \frac{1}{k}\left(\mathbb{E}\left[X^{(1)}\right] + \ldots + \mathbb{E}\left[X^{(1)}\right]\right) = \mathbb{E}\left[X^{(1)}\right] = \mathbb{E}[X] = \mathsf{popmean}\left(\{x_i\}\right)$$

so we can estimate the mean of the whole dataset from the items we see simply by averaging them.

We will not get the same value of $X^{(k)}$ each time we perform the experiment, because we see different data items in each sample. So $X^{(k)}$ has variance, and this variance is important. If it is large, then each estimate is quite different. If it is small, then the estimates cluster. Knowing the variance of $X^{(k)}$ would tell us how accurate the estimate is.

We can compute the variance easily. We write $\mathsf{popsd}\left(\{x_i\}\right)$ for the standard deviation of the whole population of x_i. Again, we write it like this to keep track of the facts that (a) it's for the whole population and (b) we don't know it.

Write $\mathbb{E}\left[(X^{(1)})^2\right]$ for the expected value of the random variable generated by drawing a single number out of the jar, squaring that number, and reporting it. Write $\mathbb{E}[X_2]$ for the expected value of the random variable generated by: drawing a number out of the jar; writing it down; returning it to the jar; then drawing a second number from the jar; and reporting the product of these two numbers. Now we have

$$\mathbb{E}\left[(X^{(1)})^2\right] = \frac{\sum_{i=1}^{N_p} x_i^2}{N_p} = \mathsf{popsd}\left(\{x_i\}\right)^2 + \mathsf{popmean}\left(\{x_i\}\right)^2$$

and

$$\mathbb{E}[X_2] = \mathbb{E}\left[X^{(1)}\right]\mathbb{E}\left[X^{(1)}\right]$$

Worked example 1.1 *Jar variances*

Show that

$$\mathbb{E}\left[(X^{(1)})^2\right] = \frac{\sum_{i=1}^{N_p} x_i^2}{N_p} = \mathsf{popsd}\left(\{x_i\}\right)^2 + \mathsf{popmean}\left(\{x_i\}\right)^2$$

Solution: First, we have $(X^{(1)})^2$ is the number obtained by taking a ticket out of the jar and squaring its data item. So

$$\mathbb{E}\left[(X^{(1)})^2\right] = \sum_{i=1}^{N_p} x_i^2 p(x_i) = \sum_{i=1}^{N_p} x_i^2 \frac{1}{N_p}.$$

Now

$$\mathsf{popsd}\left(\{x_i\}\right)^2 = \frac{\sum_{i=1}^{N_p}(x_i - \mathsf{popmean}\left(\{x_i\}\right))^2}{N_p} = \frac{\sum_{i=1}^{N_p} x_i^2}{N_p} - \mathsf{popmean}\left(\{x_i\}\right)^2$$

Worked example 1.2 *Jar variances*
Show that

$$\mathbb{E}[X_2] = \mathbb{E}\left[X^{(1)}\right]\mathbb{E}\left[X^{(1)}\right]$$

Solution: This is more interesting. It is the expected value of the random variable generated by: drawing a number out of the jar; writing it down; returning it to the jar; then drawing a second number from the jar; and reporting the product of these two numbers. Notice the two numbers are independent. Write U for the first draw, V for the second draw. Now we have

$$\begin{array}{rcl}\mathbb{E}[X_2] & = & \mathbb{E}[UV] \\ & = & \mathbb{E}[([U-\mathbb{E}[U]]+\mathbb{E}[U])([V-\mathbb{E}[V]]+\mathbb{E}[V])] \\ & = & \mathbb{E}\left[\begin{array}{c}([U-\mathbb{E}[U]])([V-\mathbb{E}[V]]) \\ +([U-\mathbb{E}[U]])\mathbb{E}[V] \\ +\mathbb{E}[U]([V-\mathbb{E}[V]]) \\ +\mathbb{E}[U]\mathbb{E}[V]\end{array}\right].\end{array}$$

But U and V are independent, so $\mathbb{E}[([U-\mathbb{E}[U]])([V-\mathbb{E}[V]])] = 0$; $\mathbb{E}[([U-\mathbb{E}[U]])\mathbb{E}[V]] = \mathbb{E}[V]\mathbb{E}[([U-\mathbb{E}[U]])] = 0$, so

$$\mathbb{E}[X_2] = \mathbb{E}[U]\mathbb{E}[V] = \mathbb{E}\left[X^{(1)}\right]\mathbb{E}\left[X^{(1)}\right].$$

Now

$$\begin{array}{rcl}\mathbb{E}\left[(X^{(k)})^2\right] & = & \frac{k\mathbb{E}\left[(X^{(1)})^2\right]+k(k-1)\mathbb{E}[X_2]}{k^2} \\ & = & \frac{\mathbb{E}\left[(X^{(1)})^2\right]+(k-1)\mathbb{E}[X_2]}{k} \\ & = & \frac{\mathsf{popsd}\left(\{x_i\}\right)^2+(k-1)\mathsf{popmean}\left(\{x_i\}\right)^2}{k}\end{array}$$

so we have

$$\begin{array}{rcl}\mathsf{Var}\left[X^{(k)}\right] & = & \mathbb{E}\left[(X^{(k)})^2\right]-\mathbb{E}\left[X^{(k)}\right]^2 \\ & = & \frac{\mathsf{popsd}\left(\{x_i\}\right)^2}{k}.\end{array}$$

This is a very useful result which is well worth remembering. There are several important consequences. First, you can estimate the mean of a dataset without seeing the whole dataset. Second, if you draw k samples, the standard deviation of your estimate of the mean is

$$\frac{\mathsf{popsd}\left(\{x_i\}\right)}{\sqrt{k}}$$

which means that (a) the more samples you draw, the better your estimate becomes and (b) the estimate improves rather slowly — for example, to halve the standard deviation in your estimate, you need to draw four times as many samples. The standard deviation of the estimate of the mean is often known as the **standard error** of the mean. This allows us to draw a helpful distinction: the population has a standard deviation, and our estimate of its mean (or other things — but we won't go into this) has a standard error.

Notice we cannot state the standard error of our estimate exactly, because we do not know $\mathsf{popsd}\left(\{x_i\}\right)$. But we could make a good estimate of $\mathsf{popsd}\left(\{x_i\}\right)$, by computing the standard deviation of the examples that we have. It is now helpful to have some notation for the particular sample we have. I will write $\sum_{i \in \text{sample}}$ for a sum over the sample items, and we will use

$$\mathsf{mean}\left(\{x_i\}\right) = \frac{\sum_{i \in \text{sample}} x_i}{k}$$

for the mean of the sample — that is, the mean of the data we actually see; this is consistent with our old notation, but there's a little reindexing to keep track of the fact we don't see all of the population. Similarly, I will write

$$\mathsf{sd}\left(\{x_i\}\right) = \sqrt{\frac{\sum_{i \in \text{sample}} (x_i - \mathsf{mean}\left(\{x_i\}\right))^2}{k}}$$

for the sample standard deviation. Again, this is the standard deviation of the data we actually see; and again, this is consistent with our old notation, again with a little reindexing to keep track of the fact we don't see all of the population. We could estimate

$$\mathsf{popsd}\left(\{x_i\}\right) \approx \mathsf{sd}\left(\{x_i\}\right)$$

and as long as we have enough examples, this estimate is good. If the number of samples is small, it is better to use

$$\mathsf{popsd}\left(\{x_i\}\right) \approx \sqrt{\frac{\sum_{i \in \text{sample}} (x_i - \mathsf{mean}\left(\{x_i\}\right))^2}{k-1}}.$$

In fact, much more is known about the distribution of $X^{(k)}$.

1.1.2 Confidence Intervals

In the previous chapter, I mentioned that adding a number of independent random variables almost always got you a normal random variable, a fact sometimes known as the central limit theorem. I didn't prove it, and I'm not going to now. But when we form $X^{(k)}$, we're adding random variables. This means that $X^{(k)}$ is a normal random variable, for sufficiently big k (for some reason, $k > 30$ is usually seen as right).

What this means is the following: Compute $X^{(k)}$ for each of a very large number of different experiments and regard the resulting numbers $e_1, \ldots, e_i, \ldots, e_r$ as data items. Convert the e_i to standard coordinates s_i, where

$$s_i = \frac{(e_i - \mathsf{mean}\left(\{e_i\}\right))}{\mathsf{std}\left(e_i\right)}$$

(i.e. by subtracting the mean of the e_i, and dividing by their standard deviation). Now construct a construct a histogram of the s. If r and k are sufficiently large, the histogram will be increasingly close to the standard normal curve.

This fact is very powerful, because it tells us how close to the true mean our estimate of the mean is likely to be. The reasoning looks like this. Our estimate of $\mathsf{popmean}\left(\{x_i\}\right)$ is $X^{(k)}$. We know that the error — which is $X^{(k)} - \mathsf{popmean}\left(\{x_i\}\right)$ — is a normal random variable, with mean 0 and standard deviation $\frac{\mathsf{popsd}(\{x_i\})}{\sqrt{k}}$. We can scale the error by the standard deviation to get

$$\hat{M} = \frac{X^{(k)} - \mathsf{popmean}\left(\{x_i\}\right)}{\left(\frac{\mathsf{popsd}(\{x_i\})}{\sqrt{k}}\right)}$$

which is a standard normal random variable. But we know rather a lot about the behaviour of standard normal random variables. In particular, from the last chapter, we have that:

- About 68% of the time, the true mean is within one standard deviation of our estimate.
- About 95% of the time, the true mean is within two standard deviations of our estimate.
- About 99% of the time, the true mean is within three standard deviations of our estimate.

We do not know the standard deviation of this normal random variable, because we do not know $\mathsf{popsd}\left(\{x_i\}\right)$ (the standard deviation of the whole population from which our samples were drawn). But we could use our estimates

$$\mathsf{popsd}\left(\{x_i\}\right) \approx \mathsf{sd}\left(\{x_i\}\right)$$

or if k is small,

$$\mathsf{popsd}\left(\{\approx\}\right)\sqrt{\frac{\sum_{i \in \text{sample}}(x_i - \mathsf{mean}\left(\{x_i\}\right))^2}{k-1}}.$$

This means that we can reason about the effect of not seeing the whole population. The most natural way to do so is to plot a **confidence interval** for our estimate. Typically, we would plot the estimate, then a set of vertical bars (which will often, but not always, be 1, 2, or 3 standard errors in length). We interpret this interval as representing the effect of sampling uncertainty on our estimate. If the jar model really did apply, then the confidence intervals have the property that the true mean lies inside the interval with probability about 0.68 (if we draw one standard error bars), 0.95 (if we draw two standard error bars) or 0.99 (if we draw three standard error bars). Figure 1.1 compares (a) numerous sample means with population means and (b) the population mean with the error bars predicted from a single sample for human height data.

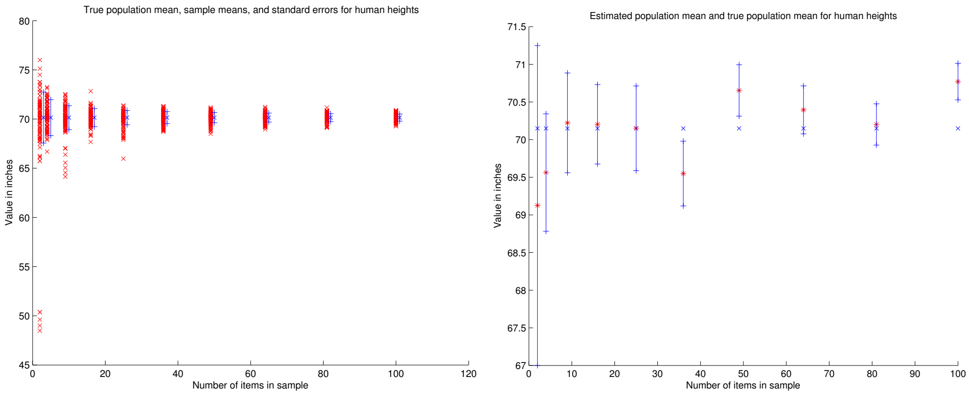

FIGURE 1.1: *I took the heights dataset (from* `http://www2.stetson.edu/~jrasp/data.htm`*; look for bodyfat.xls; outliers not removed). I then formed sampled elements with replacement to form random subsets of sizes* $(2, 4, 9, 16, \ldots, 100)$*. For each of 100 subsets of each size, I computed the sample mean — these are shown as x's on the plot on the* **left**. *I then computed the population mean, and the standard error as measured by the population standard deviation. The x to the side of each column is the population mean, and the vertical bars are one standard error above and below the population mean. Notice how (a) the sample means vary less as the sample gets bigger and (b) the sample means largely lie within the error bars. On the* **right**, *I chose one sample at random of each size; the sample mean is shown as a *. There are error bars (one standard error above and below) around the sample mean. These error bars are computed from the sample standard deviation. The population mean is the x. Notice how the population mean is within the error bars most, but not all, of the time (about 68% of the time, as they should be). The sample mean is rather a good estimate of the population mean, and the standard error is quite a reliable estimate of how well the sample mean represents the population mean.*

Now we might reasonably ask another question. We should like to know a confidence interval such that the true mean lies within it with probability p — this is equivalent to asking what is the u such that

$$\int_{-u}^{u} \frac{1}{\sqrt{2\pi}} \exp\left(-x^2/2\right) dx = p$$

(i.e. what is the range of values such that p% of standard normal random variables lies in this range). Such numbers can be extracted from the inverse of the error function (which is known as the **inverse error function**).

1.1.3 When This Model Works

In our model, there was a population of N_p data items x_i, and we saw k of them, chosen at random. In particular, each choice was fair (in the sense that each data item had the same probability of being chosen) and independent. These assumptions are very important for our analysis to apply. If our data does not have these properties, bad things can happen.

For example, assume we wish to estimate the percentage of the population that has beards. This is a mean (the data items take the value 1 for a person with a beard, and 0 without a beard). If we select people according to our model, then ask them whether they have a beard, then our estimate of the percentage of beards should behave as above.

The first thing that should strike you is that it isn't at all easy to select people according to this model. For example, we might select phone numbers at random, then call and ask the first person to answer the phone whether they have a beard; but many children won't answer the phone because they are too small. The next important problem is that errors in selecting people can lead to massive errors in your estimate. For example, imagine you decide to survey all of the people at a kindergarten on a particular day; or all of the people in a women's clothing store; or everyone attending a beard growing competition (they do exist). In each case, you will get an answer that is a very poor estimate of the right answer, and the standard error might look very small. Of course, it is easy to tell that these cases are a bad choice.

It may not be easy to tell what a good choice is. You should notice the similarity between estimating the percentage of the population that wears a beard, and estimating the percentage that will vote for a particular candidate. There is a famous example of a survey that mispredicted the result of the Dewey-Truman presidential election in 1948; poll-takers phoned random phone numbers, and asked for an opinion. But at that time, telephones tended to be owned by a small percentage of rather comfortable households, who tended to prefer one candidate, and so the polls mispredicted the result rather badly.

Sometimes, we don't really have a choice of samples. For example, we might be presented with a small dataset of (say) human body temperatures. If we can be satisfied that the people were selected rather randomly, we might be able to use this dataset to predict expected body temperature. But if we knew that the subjects had their temperatures measured because they presented themselves at the doctor with a suspected fever, then we most likely cannot use it to predict expected body temperature.

One important and valuable case where this model works is in simulation. If you can guarantee that your simulations are independent (which isn't always easy), this model applies to estimates obtained from a simulation. Notice that it is usually straightforward to build a simulation so that the i'th simulation reports an x_i where $\mathsf{popmean}\left(\{x_i\}\right)$ gives you the thing you want to measure. For example, imagine you wish to measure the probability of winning a game; then the simulation should report one when the game is won, and zero when it is lost. As another example, imagine you wish to measure the expected number of turns before a game is won; then your simulation should report the number of turns elapsed before the game was won.

1.2 HYPOTHESIS TESTING

Very often we want to draw a conclusion from data. For example, does this treatment work? Is 98.4^o the average human body temperature? Does a 20 land deck beat a 24 land deck in an MTGDAF game? and so on. Such problems can be phrased as **hypothesis tests**.

For example, imagine we hypothesize that the average human body temperature is 95^o. We collect a random sample of people, and measure their temperatures. We compute the average of these temperatures; call this $\overline{T}$. We know this average is an estimate of the mean of the original large set. We can estimate the standard error s. Our estimate may not be right, but we now know a probability distribution for the errors in the estimate that arise from sampling. In particular, we know that

$$\frac{(\overline{T} - 95^o)}{s}$$

is a standard normal random variable. Call this random variable O (for Offset). We now compute the probability that we observe a value of O at least as large, or at least as small, as the one we see *IF* the average body temperature were, in fact, 95^o. If this probability is small enough, we conclude that the average human body temperature is *not* 95^o.

There are many different kinds of hypothesis test. We just scratch the surface here. Most of the complexity occurs when the datasets are small, or when one wants to test more refined hypotheses than the ones we treat. Section **??** shows three qualitative tests for whether data is normal or not. In Section 5, I describes ways to test whether a population has a particular mean value; the details vary depending on what is assumed about the population. Section 5 tests whether two populations have the same or different means; again, some details vary according to what is assumed about the population.

Generally, quantitative hypothesis tests have the same structure. We start with a null hypothesis; we then look at data, and from it compute a number; then we determine the probability that we would have gotten a number at least as large as this by chance, if the null hypothesis were true. The resulting number is not usually thought of as a probability, because it is highly unlikely that we would be able to repeat the process of sampling and testing. Instead, it is usually thought of as representing the significance of the evidence. If the number is small, we can assert the evidence suggests the null hypothesis is untrue. The smaller the number, the more convincing the evidence that the null hypothesis is untrue.

One potential source of confusion arises. Imagine the probability that the number we observed occured by chance is high. This does not mean that we have confirmed the null hypothesis, or that the null hypothesis is true; instead, it means that we have failed to reject it. The evidence does not suggest that it is false. This is not the same as saying that the evidence suggests that it is true.

1.2.1 Is this Data Normal?

There are a variety of tests to tell whether data is normal or not. We describe only informal tests. First, a sensible thing to do is to prepare a histogram and look at it. Normal (or roughly normal) data has quite characteristic histograms, with a pronounced bump at the mean and quite light tails.

Another useful approach is the **68-95-99 rule**. A normal random variable is within one standard deviation of the mean about 68% of the time, within two standard deviations of the mean 95% of the time; and within three standard deviations of the mean about 99% of the time. You can test whether data is normal by computing the percentage of data within one, two and three standard deviations of the mean; if there is too much, or too little, it is not normal.

Finally, you could look at a **quantile-quantile plot** or a QQ plot. This is a tool that applies quite generally to comparing distributions. For this case, you standardize the data set (i.e. subtract the mean, then divide by the standard deviation). You then construct a set of target percentage values for the data. Usually, if there are N data items in the sample, one chooses $k/(N+1)$ for $k = 1, \ldots, N$. For each of these target percentage values, you compute the corresponding quantile of the data. Recall that a quantile is a value such that a given percentage of the data is below that value. These quantiles give a vector $\hat{q}_i$, with one element for each target percentage.

You then compute these quantiles for standard normal data. Call these q_i. There are two ways to do this. With appropriate manipulation of inverse error functions you can get the quantiles exactly. Alternatively, you could use a simulation — just draw a very large number of samples from a standard normal distribution, then compute their quantiles. This is less precise, but is fine for a qualitative test.

You now construct a scatter plot of $(\hat{q}_i, q_i)$ points. If the dataset is a normal dataset, then $\hat{q}_i$ should be very similar to q_i for each quantile. In turn, this means that the scatter of points should lie close to the diagonal line (Figure 1.2).

1.2.2 Doing a One Sample Test

The temperature example illustrates a recipe. We set up a null hypothesis about a population. For our purposes, this null hypothesis gives a value for an expected value. We compute a number from a dataset; this number is called a **statistic**. Call this statistic s. We then ask what is the probability that we would observe a range of values for that statistic if (a) the null hypothesis is true; and (b) the dataset is truly a random sample from the population. If this probability is small enough, we reject the null hypothesis.

We will always look at test statistics of the form

$$\frac{\text{sample mean} - \text{population mean}}{\text{standard error}}$$

because we know how these statistics are distributed. There are now a variety of one-sample tests distinguished by how one handles details in this recipe. Write $\hat{s}$ for the observed value of the statistic s. Our test could be **one-way**, where we test $P(\{s > \hat{s}\})$ (or, alternatively, $P(\{s < \hat{s}\})$). Alternatively, it could be **two-way**, where we test $P(\{s > \hat{s}\} \cup \{s < \hat{s}\})$. Generally, it is more conservative to use a two-way test, and one should do so unless there is a good reason not to. Very often, authors use one-way tests because they result in smaller p-values, and small p-values are often a requirement for publication.

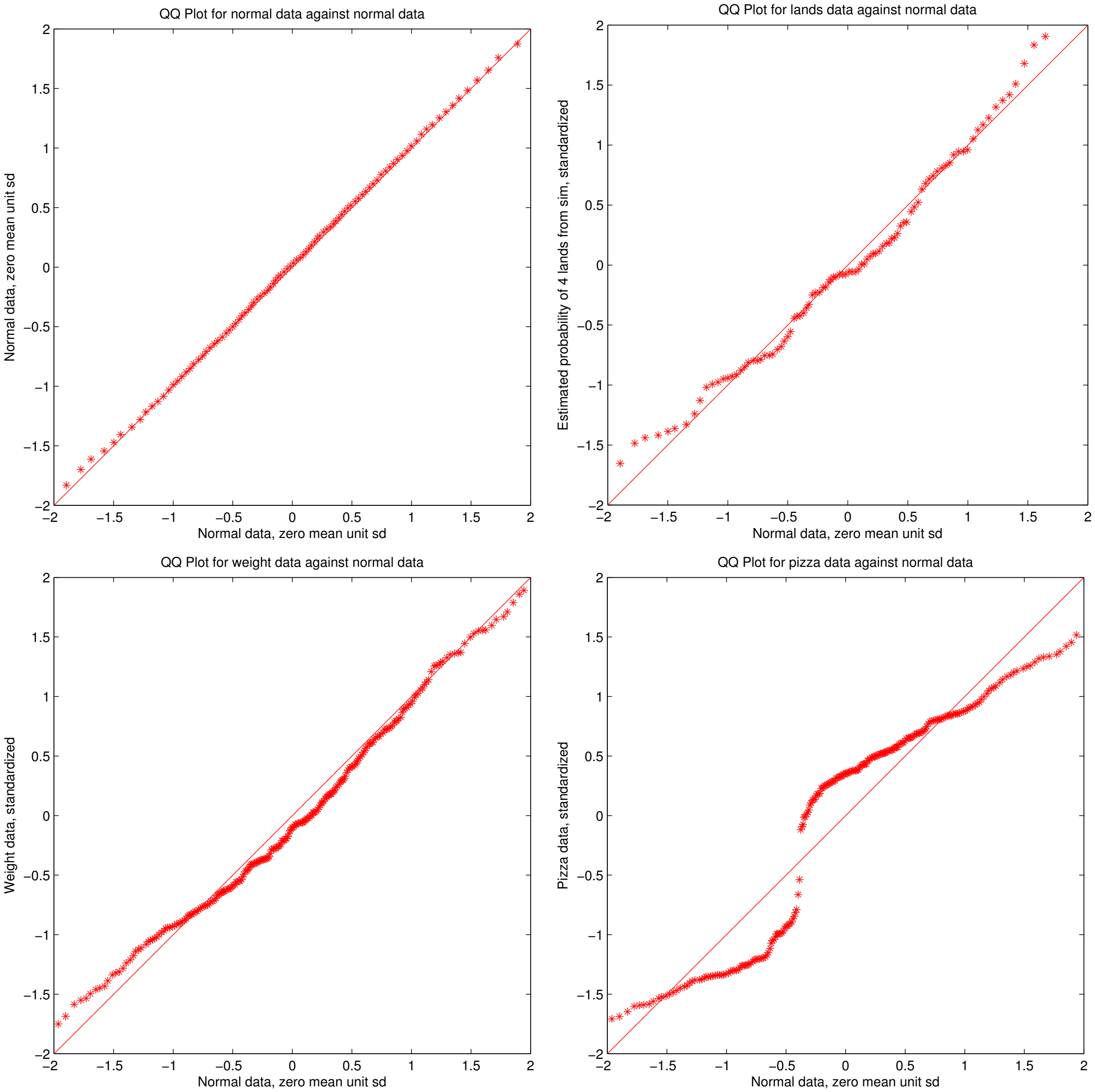

FIGURE 1.2: *On the* **top left**, *a QQ plot of normal data against normal data. Notice how the points lie close to (but not exactly on) the diagonal line. The are not on the diagonal line because these are samples of a normal distribution, rather than numbers computed exactly from that distribution. On the* **top right**, *a QQ plot of 4 land probabilities from a simulation against a normal distribution. The data is obtained by simulations of MTGDAF, estimating the probability that a player draws four lands with a hand of **** cards. Notice this data looks normal. On the* **bottom left**, *a QQ plot of the data for human weights against a normal distribution. The points lie very close to the diagonal line, suggesting this could be regarded as normal data. On the* **bottom right**, *a QQ plot of the pizza diameter data against a normal distribution. Some parts of this plot are well away from the diagonal line, and so it is likely unwise to treat this as normal data.*

The probability we compute is sometimes referred to as a **p-value**. It is conventional to reject the null hypothesis when the p-value is less than 0.05. This is sometimes called "a significance level of 5%". Sometimes, the p-value is even smaller, and this can be interpreted as very strong evidence the null hypothesis is wrong. A p-value of less than 0.01 allows one to reject the null hypothesis at "a significance level of 1%".

Z-tests: When we have a large sample, it is reasonable to assume that the sample mean is a normal random variable with mean the population mean and standard deviation given by the standard error. This means that we can compute

$$P(\{s > \hat{s}\}) = \frac{1}{\sqrt{2\pi}} \int_{\hat{s}}^{\infty} \exp\left(-x^2/2\right) dx$$

or

$$P(\{s > \hat{s}\} \cup \{s < \hat{s}\} \cup = 2 \frac{1}{\sqrt{2\pi}} \int_{\hat{s}}^{\infty} \exp\left(-x^2/2\right) dx.$$

To compute $\hat{s}$, we need to know the standard error. We estimate the standard error as above, using the sample standard deviation as an estimate of the population standard deviation. Usually, practical advice suggests that one should do this only if the sample has at least 30 elements.

T-tests: When the sample is small, the sample standard deviation is a poor estimate of the population standard deviation. The value of the sample mean that we compute minimizes the sample standard deviation, which means that the estimate tends to be a little too small. In turn, the standard error is a little too small, and there is slightly more probability that the sample mean is far from the population mean than the normal model allows for. This can be corrected for. Instead of using the standard deviation of the sample to form the standard error, we use

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

When we test, instead of using the normal distribution to compute probabilities, we use **Student's t-distribution**. This is a family of probability distributions. There are two parameters; the observed value of the statistic $\hat{s}$, and the number of degrees of freedom. The number of degrees of freedom is $k-1$ for our purposes. When the number of degrees of freedom is small, the t-distribution has rather heavier tails than the normal distribution, so the test takes into account that the standard error may be larger than we think (because the population standard deviation is larger than we expected). When the number of degrees of freedom is large, the t-distribution is very similar to the normal distribution. One can get probability (significance) values from tables, or by the Matlab function `ttest`.

1.2.3 Worked Example - Weight

Recall the height and weight data from chapter 1. We hypothesize that the average human body weight is 175lb. We assume this data set represents a random sample. It contains 252 samples. We take the average of these weights, to get 178.9lb. We know this average is an estimate of the mean of the original large set of all people. This estimate is not necessarily the mean; in fact, as we have seen, it is a normal random variable whose mean is the mean of the original data set, and whose standard deviation is given by the standard error we computed above.

We do not know the standard error exactly, because we do not know the standard deviation of the original large set. However, we can estimate it, and the estimate is quite good if we have a very large sample. Our sample is large (which usually means over 30 elements) and so we can estimate the standard error as

$$\frac{\text{standard deviation of sample}}{\sqrt{numberinsample}} = \frac{29.4}{15.9} = 1.9,$$

where the units are lb. Now our test statistic is

$$\frac{178.9 - 175}{1.9} = 2.05$$

and we know this is a normal random variable with zero mean and unit variance. We can now compute the probability that, if the average human body weight were 175lb, we would see a sample which had mean 178.9lb or greater purely by chance. This is

$$\frac{1}{\sqrt{2\pi}}\int_{2.05}^{\infty} \exp\left(\frac{-x^2}{2}\right) dx = 0.02.$$

We can interpret this as quite strong evidence that the average human body weight is not, in fact, 175lb. This probability says that, if (a) the average human body weight is 175lb and (b) we repeat the experiment (weigh 252 people and average their weights) 50 times, we would see a number as big as the one we see about once.

We could also ask what the probability is that we would see a *difference* at least as large as the one we see, if the mean weight were 175lb. This is a two-sided test. This probability, under our model, is:

$$\begin{aligned} P(\{s > 2.05\} \cup \{s < -2.05\}) &= \frac{1}{\sqrt{2\pi}}\left(\int_{2.05}^{\infty} \exp\left(\frac{-x^2}{2}\right) dx + \int_{-\infty}^{-2.05} \exp\left(\frac{-x^2}{2}\right) dx\right) \\ &= 0.041. \end{aligned}$$

We can interpret this as quite strong evidence that the average human body weight is not, in fact, 175lb. This probability says that, if (a) the average human body weight is 175lb and (b) we repeat the experiment (weigh 252 people and average their weights) 50 times, we would see a number as big as the one we see about twice.

1.2.4 Two Sample Tests

Sometimes we have two samples, and we need to know whether they come from the same, or different, populations. For example, we might observe people using two different interfaces, measure how quickly they perform a task, then ask are their performances different? As another example, we might run an application with no other applications running, and test how long it takes to run some standard tasks. Uncertainty about what the operating system, cache, etc. are up to means that this number behaves a bit like a random variable, so it is worthwhile to do this several times, yielding one set of samples. We now do this with other applications running as well, yielding another set of samples — are they different? For realistic sets of samples, the answer is always yes, because they're random samples. A better question is could the differences be the result of chance, or do these samples really come from two different populations?

One really important case occurs when it is reasonable to model both populations as normal. Then we can ask if the population means are different. This we can do with a z-test (or a t-test, if there is too little data). We set up the null hypothesis that the population means are the same. Now we know that the sample mean $\mathsf{mean}\left(\left\{x_i^{(1)}\right\}\right)$ for the first (resp. second; $\mathsf{mean}\left(\left\{x_i^{(2)}\right\}\right)$) sample has a normal distribution whose mean is the first (resp. second) population mean, and whose standard deviation is the standard error. The samples may have different sizes, so the standard errors are of different size, too.

Useful facts: *Sums and differences of normal random variables*

Let X_1 be a normal random variable with mean μ_1 and standard deviation σ_1. Let X_2 be a normal random variable with mean μ_2 and standard deviation σ_2. Let X_1 and X_2 be independent. Then we have that:

- for any constant $c_1 \neq 0$, $c_1 X_1$ is a normal random variable with mean $c_1\mu_1$ and standard deviation $c_1\sigma_1$;
- for any constant c_2, $X_1 + c_2$ is a normal random variable with mean $\mu_1 + c_2$ and standard deviation σ_1;
- X_1+X_2 is a normal random variable with mean $\mu_1+\mu_2$ and standard deviation $\sqrt{\sigma_1^2 + \sigma_2^2}$.

I will not prove these facts; we already know the expressions for means and standard deviations from our results on expectations. The only open question is to show that the sums, products, etc. are normal. This is easy for the first two results (but not worth our attention). To establish that they are normal requires a bit of integration that isn't worth our trouble; you could do reconstruct the proof from section 5's notes on sums of random variables and some work with tables.

Now we refer to the facts about normal random variables. We must have that $\mathsf{mean}\left(\left\{x_i^{(1)}\right\}\right)$ is a normal random variable, and so is $\mathsf{mean}\left(\left\{x_i^{(2)}\right\}\right)$; so this means that $\mathsf{mean}\left(\left\{x_i^{(1)}\right\}\right) - \mathsf{mean}\left(\left\{x_i^{(2)}\right\}\right)$ is also normal. Now write $\mathsf{stderr}\left(x_i^{(1)}\right)$ for the standard error of the sample mean of sample 1, etc. Then, again from our facts, the standard error of $\mathsf{mean}\left(\left\{x_i^{(1)}\right\}\right) - \mathsf{mean}\left(\left\{x_i^{(2)}\right\}\right)$ must be

$$\sqrt{\mathsf{stderr}\left(x_i^{(1)}\right)^2 + \mathsf{stderr}\left(x_i^{(2)}\right)^2}.$$

Now our test statistic is

$$\frac{\left(\left[\mathsf{mean}\left(\left\{x_i^{(1)}\right\}\right) - \mathsf{mean}\left(\left\{x_i^{(2)}\right\}\right)\right] - 0\right)}{\sqrt{\mathsf{stderr}\left(x_i^{(1)}\right)^2 + \mathsf{stderr}\left(x_i^{(2)}\right)^2}}.$$

This is a zero mean random variable with variance one. We can test this statistic with exactly the same procedures as we used for a one-sample test.

When we have a large sample, we use a z-test, so that

$$P(\{s > \hat{s}\}) = \frac{1}{\sqrt{2\pi}} \int_{\hat{s}}^{\infty} \exp\left(-x^2/2\right) dx$$

or

$$P(\{s > \hat{s}\} \cup \{s < \hat{s}\}) = 2\frac{1}{\sqrt{2\pi}} \int_{\hat{s}}^{\infty} \exp\left(-x^2/2\right) dx.$$

To compute $\hat{s}$, we need to know the standard error for each population. We estimate the standard error as above, using the sample standard deviation as an estimate of the population standard deviation. Usually, practical advice suggests that one should do this only if the sample has at least 30 elements.

When one or both of the samples are small, we should use a t-test. Instead of using the standard deviation of the samples to form the standard error, we use

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

When we test, instead of using the normal distribution to compute probabilities, we use Student's t-distribution, as above. This is a family of probability distributions. There are two parameters; the observed value of the statistic $\hat{s}$, and the number of degrees of freedom. Write s_i for the standard error of sample i and n_i for the number of elements in sample i. In this case, the number of degrees of freedom is

$$\frac{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})^2}{\frac{(\frac{s_1^2}{n_1})^2}{n_1 - 1} + \frac{(\frac{s_2^2}{n_2})^2}{n_2 - 1}}$$

a wholly non-obvious expression which I shan't derive.

1.2.5 Doing a two-sample test

We have a dataset of human temperatures (on the website). Does gender 1 have the same mean temperature as gender 2? First, we check whether the temperature data is normal with QQ plots (Figure 1.3). For gender 1, the data looks normal; for gender 2, it quite possibly is not. Nonetheless, we will treat it as normal and see what happens. Now we compute the mean temperatures and standard errors shown in the table.

Gender:	1	2
Mean:	98.10	98.39
Std Error:	0.0867	0.0922

Now the null hypothesis is that these two are the same, so the test statistic is

$$\frac{\text{difference}}{\text{std error}} = \frac{98.39 - 98.10}{\sqrt{0.0867^2 + 0.0922^2}} = 2.285$$

and we must ask what is the probability of getting a number with absolute value this big, or bigger, from a normal distribution (two-sided test). This is 0.0223. We may be able to reject the null hypothesis; alternatively, we may need to worry about the fact that gender 2 seems not to be normal. More data would likely help.

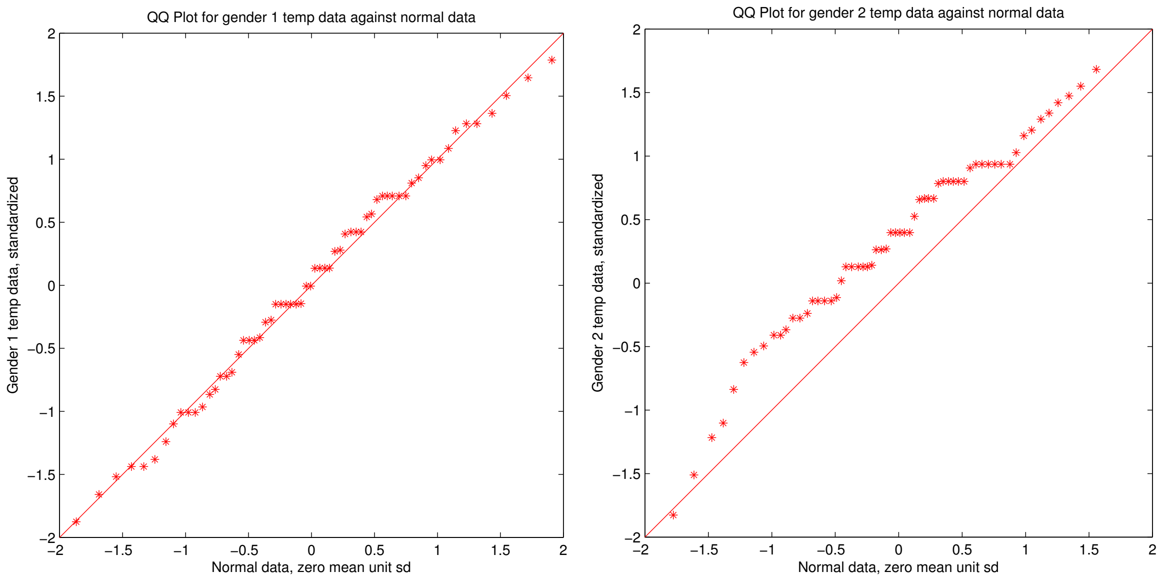

FIGURE 1.3: *On the* **left**, *a QQ plot of normal human body temperatures for gender 1, against normal data. This data is from the dataset at* `http://www2.stetson.edu/~jrasp/data.htm`*; normtemp.xls. Notice how the points lie close to (but not exactly on) the diagonal line, and look fairly normal. On the* **right**, *a QQ plot of normal body temperatures for gender 2 against a normal distribution. Notice this data looks rather less normal.*

1.2.6 Chi-squared tests

Now imagine we have a six-sided die. We throw it many times, and record which number comes up each time. We would like to know if the die is fair. It is highly unlikely that each face comes up the same number of times, even if the die is fair. Instead, there will be some variation in the frequencies observed; with what probability is that variation, or bigger, the result of chance effects?

For a case like this, where we must compare observed frequencies with theoretical frequencies, we can use a χ^2 (say "khi-squared") test. Assume we have a set of disjoint events $\mathcal{E}_1, \ldots, \mathcal{E}_k$ which cover the space of outcomes (i.e. any outcome lies in one of these events). Assume we perform k experiments, and record the number of times each event occurs. We have a null hypothesis regarding the probability of events. We can take the probability of each event and multiply by k to get a frequency under the null hypothesis. Now write $f_o(\mathcal{E}_i)$ for the observed frequency of event i; $f_t(\mathcal{E}_i)$ for the theoretical frequency of the event under the null hypothesis. We form the statistic

$$\sum_i \frac{(f_o(\mathcal{E}_i) - f_t(\mathcal{E}_i))^2}{f_t(\mathcal{E}_i)}$$

which compares the observed and actual frequency of events. It turns out that this statistic has a distribution very close to a known form, called the χ^2 distribution, as long as each count is 5 or more. The distribution has two parameters; the statistic, and the number of degrees of freedom. The number of degrees of freedom to use for a straightforward test is $k-1$; if one has to estimate a total of p parameters for the null hypothesis, this number is $k-p-1$.

After this, things follow the usual recipe. We compute the statistic; we then look at tables, or use the matlab function `chi2cdf`, to find the probability that the statistic takes this value or greater under the null hypothesis. If this is small, then we reject the null hypothesis.

Worked example 1.3 χ^2 *test for dice*

I throw a die 100 times. I record the outcomes, in the table below. Is this a fair die?

face	count
1	46
2	13
3	12
4	11
5	9
6	9

Solution: The expected frequency is 100/6 for each face. The χ^2 statistic has the value 62.7, and there are 5 degrees of freedom. We get the significance as `1-chi2cdf(62.7, 5)`, which is (basically) 3e-12. You would have to run this experiment 3e11 times to see a table as skewed as this once, by chance. The die is not fair. Notice the most important bit of this example, which is how to get the number out of matlab.

Worked example 1.4 *Is swearing Poisson?*

A famously sweary politician gives a talk. You listen to the talk, and for each of 30 intervals 1 minute long, you record the number of swearwords. You record this as a histogram (i.e. you count the number of intervals with zero swear words, with one, etc.).

no. of swear words	no. of intervals
0	13
1	9
2	8
3	5
4	5

The null hypothesis is that the politician's swearing is Poisson distributed, with intensity (λ) one. Can you reject this null hypothesis?

Solution: If the null hypothesis is true, then the probability of getting n swear words in a fixed length interval would be $\frac{\lambda^n e^{-\lambda}}{n!}$. There are 10 intervals, so the theoretical frequencies are 10 times the following probabilities

number of swear words	probability
0	0.368
1	0.368
2	0.184
3	0.061
4	0.015

so the χ^2 statistic takes the value 243.1 and there are 4 degrees of freedom. The significance `1-chi2cdf(243.1, 4)` is indistinguishable from zero by matlab, so you can firmly reject the null hypothesis. Of course, the intensity may be wrong; but we don't know how to deal with that, and will have to discuss that question in the next chapter.

Worked example 1.5 *Is gender 2 temperature normal?*

Recall we used body temperature data for two genders in earlier examples. The gender 2 data looked as though it might not be normal. Use a χ^2 test to tell whether it is normal with mean 98.4^o and standard deviation 0.743 or not.

Solution: The null hypothesis is that the data is normal with mean 98.4^o and standard deviation 0.743. We break up the range into five buckets (less than $97.65 = 98.4 - 0.74$; between 97.65 and $98 = 98.4 - 0.743/2$; between 98 and $98.76 = 98.4 + 0.743/2$; and greater than 99.14). With a little work with error functions, we get that the theoretical frequency in each bucket under the null hypothesis is $(10.3126; 9.7423; 24.8901; 9.7423; 10.3126)$. The actual frequencies are $(7; 13; 26; 12; 7)$. The χ^2 statistic is about 4e3, and the significance level is essentially zero. This data isn't normal with the parameters given (though it might be normal with a different mean and a different standard deviation). Looking at the frequencies suggests the problem; there are far too few temperatures far away from the mean for this to be normal, and far too many in the center bucket.

1.2.7 Obiter Dicta

Keep in mind that a p-value is not revealed truth. It is a summary of the evidence against a null hypothesis. It tells you the probability that you would see the number you see, or worse, if the null hypothesis were true. There are parts of the scientific world — rather depraved parts, I should add — where one cannot get papers published unless something in the papers (a) has a p-value and (b) has a p-value below 0.05. Substituting ritual for thought is never wise. Among other things, it leads to a variety of tricks for finding things that have small p-values. This sort of behavior should be disdained.