Samples and Populations

6

Very often the data we see is a small part of the data we could have seen. The data we could have observed, if we could have seen everything, is the **population**. I will write populations like random variables with capital letters to emphasize we don't actually know the whole population. The data we actually have is the **sample** (lower case, as usual). We would like to know the **population mean**, which we write $\mathsf{popmean}(\{X\})$. We must estimate this using the sample.

This situation occurs very often. For example, imagine we wish to know the average weight of a rat. This isn't random; you could weigh every rat on the planet, and then average the answers. But doing so would absurd (among other things, you'd have to weigh them all at the same time, which would be tricky). Instead, we weigh a small set of rats, chosen at random but rather carefully so. If we have chosen sufficiently carefully, then we can say a great deal from the sample alone.

6.1 The Sample Mean

Assume we have a population $\{X\}$, for $i = 1, \ldots, N_p$. Notice the subscript here—this is the number of items in the population. The population could be unreasonably big: for example, it could consist of all the people in the world. We want to know the mean of this population, but we do not get to see the whole thing. Instead, we see a sample.

How the sample is obtained is key to describing the population. We will focus on only one model (there are lots of others). In our model, the sample is obtained by choosing a fixed number of data items. Write N for the number of data items in the sample. I use N to remind you of the size of a dataset, because most datasets are samples. We expect N is a lot smaller than N_p. Each item is chosen independently, and fairly. This means 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 natural way to think about sampling with replacement is to imagine the data items as being written on tickets, which are placed in an urn (old-fashioned word for a jar, now used mainly by statisticians and morticians). You obtain the sample by repeating the following experiment N times: shake the urn; take a ticket from the urn and write down the data on the ticket; put it back in the urn. Notice that, in this case, each sample is drawn from the same urn. This is important, and makes the analysis easier. If we had not put the ticket back, the urn would change between samples.

6.1.1 The Sample Mean Is an Estimate of the Population Mean

We would like to estimate the mean of the whole dataset from the items that we actually see. Imagine we draw N tickets from the urn as above, and average the values. The result is a random variable, because different draws of N tickets will give us different values. Write $X^{(N)}$ for this random variable, which is referred to as the **sample mean**. Because expectations are linear, we must have that

© Springer International Publishing AG 2018
D. Forsyth, *Probability and Statistics for Computer Science*,
https://doi.org/10.1007/978-3-319-64410-3_6

$$\mathbb{E}\left[X^{(N)}\right] = \frac{1}{N}\left(\mathbb{E}\left[X^{(1)}\right] + \ldots + \mathbb{E}\left[X^{(1)}\right]\right) = \mathbb{E}\left[X^{(1)}\right]$$

(where $X^{(1)}$ is the random variable whose value is obtained by drawing one ticket from the urn). Now

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

which is the mean value of the items in the urn. This means that

$$\mathbb{E}\left[X^{(N)}\right] = \mathsf{popmean}\left(\{X\}\right).$$

Under our sampling model, the expected value of the sample mean is the population mean.

Useful Facts 6.1 (Properties of Sample and Population Means)
The sample mean is a random variable. It is random, because different samples from the population will have different values of the sample mean. The expected value of this random variable is the population mean.

6.1.2 The Variance of the Sample Mean

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

We write $\mathsf{popsd}\left(\{X\}\right)$ for the standard deviation of the whole population $\{X\}$. 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—and usually can't—know it. We can compute the variance of $X^{(N)}$ (the sample mean) easily. We have

$$\mathsf{var}\left[X^{(N)}\right] = \mathbb{E}\left[(X^{(N)})^2\right] - \mathbb{E}\left[X^{(N)}\right]^2 = \mathbb{E}\left[(X^{(N)})^2\right] - \left(\mathsf{popmean}\left(\{X\}\right)\right)^2$$

so we need to know $\mathbb{E}\left[(X^{(N)})^2\right]$. We can compute this by writing

$$X^{(N)} = \frac{1}{N}(X_1 + X_2 + \ldots X_N)$$

where X_1 is the value of the first ticket drawn from the urn, etc. We then have

$$X^{(N)2} = \left(\frac{1}{N}\right)^2 \left(\begin{array}{l} X^2{}_1 + X^2{}_2 + \ldots X^2{}_N + X_1 X_2 + \ldots \\ X_1 X_k + X_2 X_1 + \ldots X_2 X_N + \ldots X_{N-1} X_N \end{array} \right).$$

Expectations are linear, so we have that

$$\mathbb{E}\left[(X^{(N)})^2\right] = \left(\frac{1}{N}\right)^2 \left(\begin{array}{l} \mathbb{E}\left[{X^2}_1\right] + \mathbb{E}\left[{X^2}_2\right] + \ldots \mathbb{E}\left[{X^2}_N\right] + \mathbb{E}[X_1X_2] + \\ \ldots \mathbb{E}[X_1X_N] + \mathbb{E}[X_2X_1] + \ldots \mathbb{E}[X_{N-1}X_N] \end{array} \right).$$

The *order* in which the tickets are drawn from the urn doesn't matter, because each time we draw a ticket we draw from the same urn. This means that $\mathbb{E}\left[{X^2}_1\right] = \mathbb{E}\left[{X^2}_2\right] = \ldots = \mathbb{E}\left[{X^2}_N\right]$. You can think of this term as the expected value of the random variable generated by: drawing a single number out of the urn; squaring that number; and reporting the square. Notice that $\mathbb{E}\left[{X^2}_1\right] = \mathbb{E}\left[(X^{(1)})^2\right]$ (look at the definition of $X^{(1)}$).

Because the *order* doesn't matter, we also have that $\mathbb{E}[X_1X_2] = \mathbb{E}[X_1X_3] = \ldots \mathbb{E}[X_{N-1}X_N]$. You can think of this term as the expected value of the random variable generated by: drawing a number out of the urn; writing it down; returning it to the urn; then drawing a second number from the urn; and reporting the product of these two numbers. So we can write

$$\mathbb{E}\left[{X^{(N)}}^2\right] = \left(\frac{1}{N}\right)^2 \left(N\mathbb{E}\left[(X^{(1)})^2\right] + N(N-1)\mathbb{E}[X_1X_2]\right)$$

and these two terms are quite easy to evaluate.

Worked example 6.1 (Urn Variances) Show that

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

Solution First, we have $(X^{(1)})^2$ is the number obtained by taking a ticket out of the urn uniformly and at random and squaring its data item. Now

$$\begin{aligned} \mathsf{popsd}\left(\{X\}\right)^2 &= \mathbb{E}\left[(X^{(1)})^2\right] - \mathbb{E}\left[X^{(1)}\right]^2 \\ &= \mathbb{E}\left[(X^{(1)})^2\right] - \mathsf{popmean}(\{X\})^2 \end{aligned}$$

so

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

Worked example 6.2 (Urn Variances) Show that

$$\mathbb{E}[X_1X_2] = \mathsf{popmean}\left(\{X\}\right)^2$$

Solution This looks hard, but isn't. Recall from the facts in Chap. 4 (useful facts 4.6, page 97) that if X and Y are independent random variables, $\mathbb{E}[XY] = \mathbb{E}[X]\mathbb{E}[Y]$. But X_1 and X_2 are independent—they are different random draws from the same urn. So

$$\mathbb{E}[X_1X_2] = \mathbb{E}[X_1]\mathbb{E}[X_2]$$

but $\mathbb{E}[X_1] = \mathbb{E}[X_2]$ (they are draws from the same urn) and $\mathbb{E}[X] = \mathsf{popmean}\left(\{X\}\right)$. So

$$\mathbb{E}[X_1X_2] = \mathsf{popmean}\left(\{X\}\right)^2.$$

Now

$$\begin{aligned} \mathbb{E}\left[(X^{(N)})^2\right] &= \frac{N\mathbb{E}\left[(X^{(1)})^2\right] + N(N-1)\mathbb{E}[X_1X_2]}{N^2} \\ &= \frac{\mathbb{E}\left[(X^{(1)})^2\right] + (N-1)\mathbb{E}[X_1X_2]}{N} \end{aligned}$$

$$= \frac{(\text{popsd}(\{X\})^2 + \text{popmean}(\{X\})^2) + (N-1)\text{popmean}(\{X\})^2}{N}$$

$$= \frac{\text{popsd}(\{X\})^2}{N} + \text{popmean}(\{X\})^2$$

so we have

$$\begin{aligned}\text{var}\left[X^{(N)}\right] &= \mathbb{E}\left[(X^{(N)})^2\right] - \mathbb{E}\left[X^{(N)}\right]^2 \\ &= \frac{\text{popsd}(\{X\})^2}{N} + \text{popmean}(\{X\})^2 - \text{popmean}(\{X\})^2 \\ &= \frac{\text{popsd}(\{X\})^2}{N}.\end{aligned}$$

This is a very useful result which is well worth remembering together with our facts on the sample mean, so we'll put them in a box together.

Useful Facts 6.2 (Expressions for Mean and Variance of the Sample Mean)
The sample mean is a random variable. Write $X^{(N)}$ for the mean of N samples. We have that:

$$\mathbb{E}\left[X^{(N)}\right] = \text{popmean}(\{X\})$$

$$\text{var}\left[X^{(N)}\right] = \frac{\text{popsd}(\{X\})^2}{N}$$

$$\text{std}\left(X^{(N)}\right) = \frac{\text{popsd}(\{X\})}{\sqrt{N}}$$

The consequence is this: If you draw N samples, the standard deviation of your estimate of the mean is

$$\frac{\text{popsd}(\{X\})}{\sqrt{N}}$$

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.

6.1.3 When The Urn Model Works

In our model, there was a population of N_p data items x_i, and we saw N 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 deviation of this estimate 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 without a lot of extra work.

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 $\mathsf{popmean}(\{X\})$ 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.

6.1.4 Distributions Are Like Populations

Up to now, we have assumed that there is a large population of data items from which we drew a sample. The sample was drawn from the population uniformly at random, and with replacement. We used this sample to reason about the population. But the ideas depended on the population only to the extent that (a) the population is very big; (b) there was a population mean; and (c) the population was sampled uniformly at random, and with replacement. This suggests that we can replace the population with a probability distribution and the sampling process with drawing IID samples from the population.

Now imagine that we have a set of N data items x_i drawn as IID samples from some distribution $P(X)$. We require that the mean and variance of this distribution exist (there are some distributions for which this criterion does not apply; they're of no interest to us). The derivations of Sects. 6.1.1 and 6.1.2 work fine for this case. We have that

$$X^{(N)} = \frac{\sum_i x_i}{N}$$

is a random variable, because different sets of IID samples will have different values. We will have that

$$\mathbb{E}\left[X^{(N)}\right] = \mathbb{E}_{P(X)}[X]$$

(i.e. the expected value of $X^{(N)}$ will be the mean of $P(X)$) and that

$$\mathsf{var}\left[X^{(N)}\right] = \frac{\mathsf{var}[P(X)]}{N}$$

(i.e. the variance of the estimate of the mean is the variance of $P(X)$ divided by N). It's important to keep track of the difference between the variance of the estimate of the mean—which describes how estimates of the mean from different samples will differ—and the variance of the original probability distribution.

6.2 Confidence Intervals

It can be important to know what range a parameter could take, and still be consistent with the data. This is particularly true when there are safety or legal considerations to worry about. Imagine you have a machine that fills cereal boxes. Each box gets a quantity of cereal that is random, but has low variance. If the weight of cereal in any box is below the amount printed on the label, you might be in trouble. When you choose the amount to print on the label, estimating the mean weight of cereal as a number might not be particularly helpful. If that estimate is a little low, you could have problems. Instead, what you'd like to know is an interval that the mean lies in with very high probability. Then you can print a label number that is smaller than the smallest in the interval, and be confident that the amount in the box is more than the amount on the label.

6.2.1 Constructing Confidence Intervals

A **statistic** is a function of a dataset. One example of a statistic is the mean of dataset. You should notice that you can write out this function without actually drawing a sample. We observe the value of a statistic by applying the function to the dataset we obtained by drawing a sample. The dataset is random, because it is either a sample from a population or an IID sample from a distribution model. This means that we should think of the value of the statistic as the observed value of a random variable—if we had a different sample from the same population (resp. IID sample from the same distribution) we would compute a different value of the statistic. We are interested in the expected value of this random variable, rather than the observed value. We would like to use the observed value of the statistic to construct an interval where we have a specified confidence that the expected value lies in the interval.

The meaning of these intervals can be somewhat delicate (and so can the constructions). I will show how to build these intervals in the case of a population mean. Here we wish to use the value of the sample mean to construct an interval for the population mean. We need to be careful about the meaning of the interval, because there is nothing random once the sample has been drawn. We can't talk sensibly about the probability that the population mean lies inside the interval, because the population mean isn't random. The interval will depend on the value of the sample mean, and so will differ from sample to sample. Choose some fraction f; we will construct an interval so that, for that fraction of samples, the population mean will lie inside the interval constructed from the sample mean. The construction requires some detailed study of the distribution of sample means.

Definition 6.1 (Confidence Interval for a Population Mean) Choose some fraction f. An f confidence interval for a population mean is an interval constructed using the sample mean. It has the property that for that fraction f of all samples, the population mean will lie inside the interval constructed from each sample's mean.

Definition 6.2 (Centered Confidence Interval for a Population Mean) Choose some $0 < \alpha < 0.5$. A $1 - 2\alpha$ centered confidence interval for a population mean is an interval $[a, b]$ constructed using the sample mean. It has the property that for α of all samples, the population mean is greater than b, and for another α of all samples, the population mean is less than a. For all other samples, the population mean will lie inside the interval.

6.2.2 Estimating the Variance of the Sample Mean

Recall the variance of the sample mean is

$$\frac{\mathsf{popsd}\left(\{X\}\right)^2}{N}$$

which isn't much help currently, because we do not know $\mathsf{popsd}\left(\{X\}\right)$. But we might estimate $\mathsf{popsd}\left(\{X\}\right)$ by computing the standard deviation of the examples that we have. I will write the sample using the notation for datasets, and will use

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

for the mean of the sample—that is, the mean of the data we actually see. Similarly, I will write

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

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. We could estimate

$$\mathsf{popsd}\left(\{X\}\right) \approx \mathsf{std}\left(\{x\}\right)$$

and as long as we have enough examples, this estimate is good. It turns out that, if the number of samples N is small, it is better to use

$$\mathsf{popsd}\left(\{X\}\right) \approx \sqrt{\frac{\sum_i (x_i - \mathsf{mean}\left(\{x\}\right))^2}{N-1}}.$$

The exercises show that

$$\mathbb{E}[\mathsf{popsd}\left(\{X\}\right)] = \mathsf{std}\left(x\right) \sqrt{\left(\frac{N}{N-1}\right)}.$$

This means that estimating $\mathsf{popsd}\left(\{X\}\right)$ using $\mathsf{std}\left(x\right)$ will produce a number that is reliably slightly too small. The estimate is referred to as a **biased estimate**, because the expected value of the estimate is not what we want it to be. In this case, it is straightforward to produce an **unbiased estimate**, and an unbiased estimate of $\mathsf{popsd}\left(\{X\}\right)$ is

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

The standard deviation of the estimate of the mean is often known as the **standard error** of the mean. I will write

$$\mathsf{stderr}\left(\{x\}\right) = \frac{\mathsf{stdunbiased}\left(\{x\}\right)}{\sqrt{N}}.$$

This term allows us to draw a helpful distinction: the population has a standard deviation, and our estimate of its mean has a standard error.

Definition 6.3 (Standard Error) Write $X^{(N)}$ for the mean of N samples x_i. $X^{(N)}$ is a random variable. An estimate of the standard deviation of $X^{(N)}$ is

$$\frac{\mathsf{stdunbiased}\left(\{x\}\right)}{\sqrt{N}}.$$

This estimate is the standard error of the mean.

Here is what is causing the bias in our estimate of σ^2. The numerator of S^2 is a sum of N numbers, but these numbers are not independent, because

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

This means that there are only $N-1$ independent numbers. Another way to see this is that, if you have $N-1$ of the terms in the sum, you can infer the N'th; in turn, counting the N'th number in the mean is unwise. Statisticians say that this average has $N-1$ **degrees of freedom**.

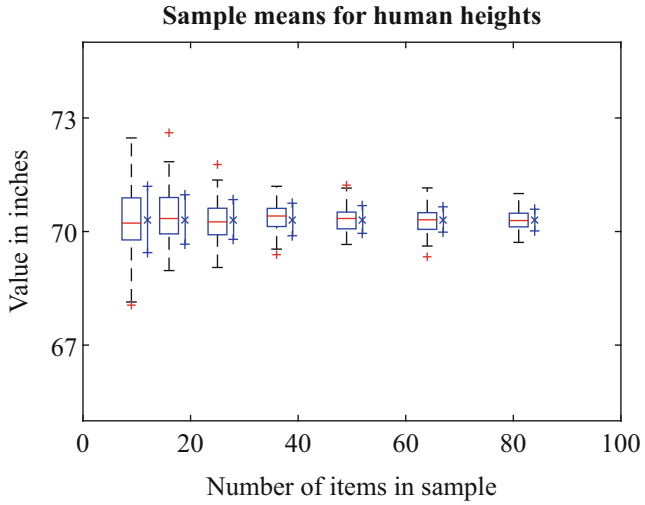

Fig. 6.1 A simple demonstration that sample means behave as described, by computing sample means from the heights dataset. I sampled elements with replacement to form random subsets of sizes $(9, 16, \ldots, 81)$. For each of 100 subsets of each size, I computed the sample mean. This means that there are 100 sample means for each sample size. I have represented these means by a boxplot. 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, as they should

Worked example 6.3 (Simulations Confirm the Standard Error Estimate) Compare the standard error of a mean estimate with the standard deviation predicted using the population from which the sample was drawn.

Solution I used the heights column from the bodyfat dataset (from http://www2.stetson.edu/~jrasp/data.htm; look for bodyfat.xls). I removed the single height outlier. I simulated the population using the whole dataset (251 items), then drew numerous samples of various sizes, with replacement. I computed the mean of each of these sets of samples. Figure 6.1 shows a scatter plot of sample means for different samples, using a set of sizes $(9, 16, \ldots, 81)$. I have also plotted the population mean, and the true 1-standard error bars (i.e. using the population standard deviation) for each of these sample sizes. Notice how most sample means lie within the 1-standard error bars, as they should.

6.2.3 The Probability Distribution of the Sample Mean

The sample mean is a random variable. We know an expression for its mean and for its variance in terms of population mean and variance, and we know that for sufficiently large samples the sample mean is a normal random variable. We have that

$$\frac{\mathsf{mean}\,(\{x\}) - \mathsf{popmean}\,(\{X\})}{\mathsf{popsd}\,(\{X\})/\sqrt{N}}$$

is a standard normal random variable. But we have to *estimate* the variance of the sample mean, and that estimate will be slightly wrong. Recall the notation

$$\mathsf{stderr}\,(\{x\}) = \frac{\mathsf{stdunbiased}\,(\{x\})}{\sqrt{N}}.$$

We are interested in the distribution of

$$T = \frac{\mathsf{mean}\,(\{x\}) - \mathsf{popmean}\,(\{X\})}{\mathsf{stderr}\,(\{x\})}$$

which is a random variable, because the samples are random.

When N is small, the estimate of the population standard deviation using the sample standard deviation is more likely to be smaller than it should be, because there some probability of choosing a sample whose variance is smaller than the population's variance. In turn, the distance between the population mean and the sample mean in standard error units may be larger than a normal distribution predicts. This means the distribution of T must depend on N. It does so through the number of degrees of freedom in the estimate of the variance of the sample mean (which is $N - 1$). This means that there is a *family* of distributions for T, indexed by the number of degrees of freedom. This family is of known form, and is known as **t-distribution**. A random variable whose distribution is a t-distribution is often known as a **t-random variable**. You will often see this referred to as Student's t-distribution, after the inventor who wrote very important statistical papers under a pseudonym because he was concerned his employer wouldn't like them (the story is worth looking up).

When the number of degrees of freedom is small, the t-distribution has rather heavier tails than the normal distribution. However, when the number of degrees of freedom is large, the t-distribution is very similar to the normal distribution. If N is large (for some reason, 30 seems to be the magic number), then it is usually safe to regard the t-distribution as being the same as a normal distribution.

Definition 6.4 (T-Distribution) Student's t-distribution is a probability distribution taken from a family, indexed by a number (the degrees of freedom of the distribution). The form of the distribution is not important to us. We will obtain values from tables or from software, and typically only need values of the cumulative distribution. When the number of degrees of freedom is large, the distribution is very similar to a normal distribution; otherwise, the tails are somewhat heavier than those of a normal distribution.

Definition 6.5 (T-Random Variable) A t-random variable is a random variable whose distribution is a Student's t-distribution.

Remember this: *The sample mean yields the value of a t random variable. In particular,*

$$T = \frac{\mathsf{mean}\left(\{x\}\right) - \mathsf{popmean}\left(\{X\}\right)}{\mathsf{stderr}\left(\{x\}\right)}$$

has a t-distribution with $N - 1$ degrees of freedom.

Remember this: *If N is large enough, the sample mean yields the value of a standard normal random variable. In particular, if N is large enough,*

$$Z = \frac{\mathsf{mean}\left(\{x\}\right) - \mathsf{popmean}\left(\{X\}\right)}{\mathsf{stderr}\left(\{x\}\right)}$$

is a standard normal random variable.

6.2.4 Confidence Intervals for Population Means

Here is a construction for a confidence interval for the mean of a population. Draw a sample randomly and with replacement of N items (write $\{x\}$ for the sample), and compute the sample mean. The sample mean is the value of a random variable—random, because it depends on the randomly drawn sample—whose probability distribution we know. Our estimate of the unknown number $\mathsf{popmean}\left(\{X\}\right)$ is the mean of the sample we have, which we write $\mathsf{mean}\left(\{x\}\right)$. We know that

$$T = \frac{\mathsf{mean}\left(\{x\}\right) - \mathsf{popmean}\left(\{X\}\right)}{\mathsf{stderr}\left(\{x\}\right)}$$

has a t-distribution. Now assume that N is large, so that the t-distribution is very similar to a standard normal distribution. But we know rather a lot about the behaviour of standard normal random variables. For about 68% of samples, t (the value of T) will lie between -1 and 1, and so on. In turn, this means that for about 68% of samples, $\mathsf{popmean}(\{X\})$ will lie in the interval between $\mathsf{mean}(\{x\}) - \mathsf{stderr}(\{x\})$ and $\mathsf{mean}(\{x\}) + \mathsf{stderr}(\{x\})$, and so on.

Useful Facts 6.3 (Easy Confidence Intervals for a Big Sample)
Assume the sample is large enough so that $\mathsf{mean}(\{x\}) - \mathsf{popmean}(\{X\})/\mathsf{stderr}(\{x\})$ is a standard normal random variable. Recall the facts in box 5.10 (page 125). These yield
For about 68% of samples:

$$\mathsf{mean}(\{x\}) - \mathsf{stderr}(\{x\}) \leq \mathsf{popmean}(\{X\}) \leq \mathsf{mean}(\{x\}) + \mathsf{stderr}(\{x\}).$$

For about 95% of samples:

$$\mathsf{mean}(\{x\}) - 2\mathsf{stderr}(\{x\}) \leq \mathsf{popmean}(\{X\}) \leq \mathsf{mean}(\{x\}) + 2\mathsf{stderr}(\{x\}).$$

For about 99% of samples:

$$\mathsf{mean}(\{x\}) - 3\mathsf{stderr}(\{x\}) \leq \mathsf{popmean}(\{X\}) \leq \mathsf{mean}(\{x\}) + 3\mathsf{stderr}(\{x\}).$$

Worked example 6.4 (The Weight of Female Mice Eating Chow) Give a 95% confidence interval for the weight of a female mouse who ate chow, based on the dataset at http://cgd.jax.org/datasets/phenotype/SvensonDO.shtml.

Solution There are great datasets dealing with a wide range of genotype and phenotype variations in mice at this URL. The one to look at is Churchill.Mamm.Gen.2012.phenotypes.csv, which has information about 150 mice. 100 were fed with chow, and 50 with a high fat diet. You should look at Weight2, which seems to be a weight around the time the mouse was sacrificed. If we focus on the female mice who ate chow and whose weights are available (48 by my count), we find a mean weight of 27.78 gr, and a standard error of 0.70 gr (remember to divide by the square root of 48). This means that the interval we want runs from 26.38 to 29.18 gr.

The authors ask that anyone using this data should cite the papers: *High-Resolution Genetic Mapping Using the Mouse Diversity Outbred Population*, Svenson KL, Gatti DM, Valdar W, Welsh CE, Cheng R, Chesler EJ, Palmer AA, McMillan L, Churchill GA. **Genetics**. 2012 Feb;190(2):437–47; and *The Diversity Outbred Mouse Population* Churchill GA, Gatti DM, Munger SC, Svenson KL **Mammalian Genome 2012**, Aug 15.

We can plot the confidence interval by drawing **error bars**—draw a bar one (or two, or three) standard errors up and down from the estimate. We interpret this interval as representing the effect of sampling uncertainty on our estimate. If the urn model really did apply, then the confidence intervals have the property that the true mean lies inside the interval for about 68% of possible samples (for one standard error bars; or 95% for two; etc.).

Procedure 6.1 (Constructing a Centered $1-2\alpha$ Confidence Interval for a Population Mean for a Large Sample)
Draw a sample $\{x\}$ of N items from a population. Recall

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

Estimate the standard error using

$$\mathsf{stderr}(\{x\}) = \frac{\mathsf{stdunbiased}(\{x\})}{\sqrt{N}}.$$

(continued)

If N is large enough, the variable

$$T = \frac{\text{mean}(\{x\}) - \text{popmean}(\{X\})}{\text{stderr}(\{x\})}$$

is a standard normal random variable.

Compute b such that for a standard normal random variable, $P(\{T \geq b\}) = \alpha$. You can do this using tables or software. The confidence interval is then

$$[\text{mean}(\{x\}) - b \times \text{stderr}(\{x\}), \\ \text{mean}(\{x\}) + b \times \text{stderr}(\{x\})].$$

Now assume that N is small enough so that T is a t-random variable (which will have $N - 1$ degrees of freedom). Assume we wish to have a centered, $1 - 2\alpha$ confidence interval. We can use tables or software to choose a value a so that $P(\{T \leq a\}) = \alpha$ and a value b such that $P(\{T \geq b\}) = \alpha$. In fact, we will have $a = -b$. This is because t-random variables have the property that $P(\{T \geq b\}) = P(\{T \leq -b\})$ (just like standard normal random variables; if you're in doubt, check this point). Then for $1 - 2\alpha$ of all samples, we have

$$-b \leq T \leq b.$$

This means that for $1 - 2\alpha$ of all samples,

$$\begin{aligned}\text{mean}(\{x\}) - b \times \text{stderr}(\{x\}) &\leq \text{popmean}(\{X\}) \\ &\leq \text{mean}(\{x\}) + b \times \text{stderr}(\{x\})\end{aligned}$$

and so we have a centered, $1 - 2\alpha$ confidence interval.

Procedure 6.2 (Constructing a Centered $1 - 2\alpha$ Confidence Interval for a Population Mean for a Small Sample) Draw a sample $\{x\}$ of N items from a population. Recall

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

Estimate the standard error using

$$\text{stderr}(\{x\}) = \frac{\text{stdunbiased}(\{x\})}{\sqrt{N}}.$$

If N is small, the variable

$$T = \frac{\text{mean}(\{x\}) - \text{popmean}(\{X\})}{\text{stderr}(\{x\})}$$

is a t-random variable.

Compute b such that for a t-random variable, $P(\{T \geq b\}) = \alpha$. You can do this using tables or software. The confidence interval is then

$$[\text{mean}(\{x\}) - b \times \text{stderr}(\{x\}), \\ \text{mean}(\{x\}) + b \times \text{stderr}(\{x\})].$$

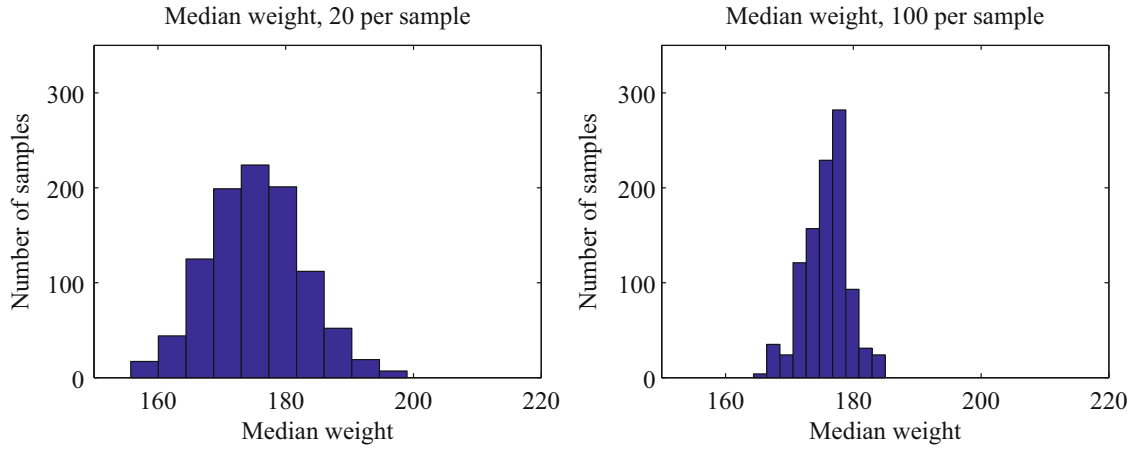

Fig. 6.2 I took the weights dataset used all 253 measurements to represent a population. Rather than compute the median of the whole population, I chose to compute the median of a randomly chosen sample. The figures show a histogram of 1000 different values of the median, computed for 1000 different samples (of size 20 on the *left*, and of size 100 on the *right*). Notice that (a) there is a moderate amount of variation in the median of the sample; (b) these histograms look normal, and appear to have about the same mean; (c) increasing the size of the sample has reduced the spread of the histogram

6.2.5 Standard Error Estimates from Simulation

We were able to produce convenient and useful estimates of standard error for sample means. But what happens if we want to reason about, say, the median of a population? Estimating the standard error of a median is difficult mathematically, and estimating the standard error of other interesting statistics can be difficult, too. This is an important problem, because our methods for building confidence intervals and for testing hypotheses rely on being able to construct standard error estimates. Quite simple simulation methods give very good estimates of standard error.

The distribution of median values for different samples of a population looks normal by simple tests. For Fig. 6.2, I assumed that all 253 weight measurements represented the entire population, then simulated what would happen for different random samples (with replacement) of different sizes. Figure 6.2 suggests that the sample median behaves quite like the sample mean as the random sample changes. Different samples have different medians, but the distribution of values looks fairly normal. When there are more elements in the sample, the standard deviation of median values is smaller, but we have no expression for this standard deviation in terms of the sample.

There is a method, known as the **bootstrap**, which gives a very good estimate of the standard error of any statistic. Assume we wish to estimate the standard error of a statistic $S(\{x\})$, which is a function of our dataset $\{x\}$ of N data items. We compute r **bootstrap replicates** of this sample. Each replicate is obtained by sampling the dataset uniformly, and with replacement. One helpful way to think of this is that we are modelling our dataset as a sample of a probability distribution. This distribution, sometimes known as the **empirical distribution**, has probability $1/N$ at each of the data items we see, and zero elsewhere. Now to obtain replicates, we simply draw new sets of IID samples from this probability distribution. Notice that the bootstrap replicates are *not* a random permutation of the dataset; instead, we select one data item fairly and at random from the whole dataset N times. This means we expect a particular bootstrap replicate will have multiple copies of some data items, and no copies of others.

We write $\{x\}_i$ for the i'th bootstrap replicate of the dataset. We now compute

$$\overline{S} = \frac{\sum_i S(\{x\}_i)}{r}$$

and the standard error estimate for S is given by:

$$\text{stderr}\left(\{S\}\right) = \sqrt{\frac{\sum_i \left[S(\{x\}_i) - \overline{S}\right]^2}{r-1}}$$

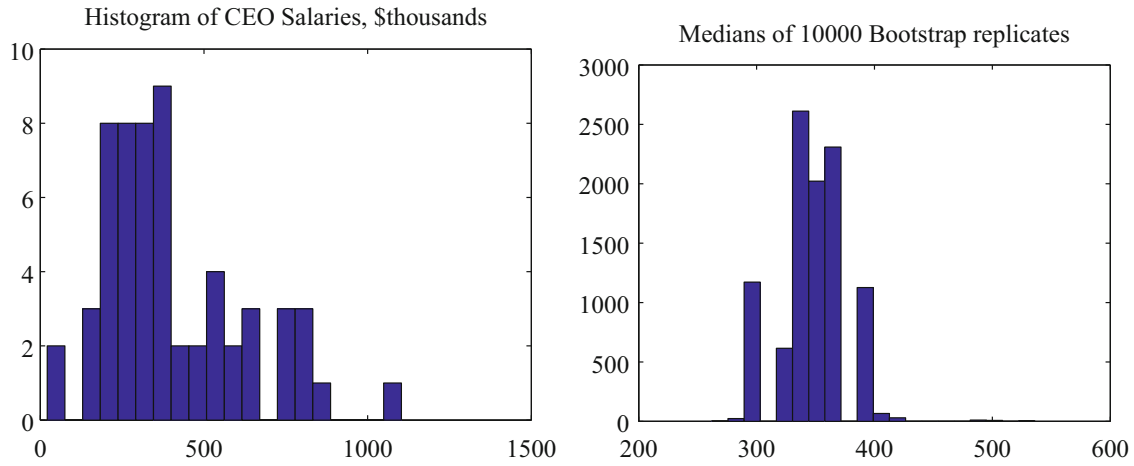

Fig. 6.3 On the *left*, a histogram of salaries for CEO's of small companies in 1993, from the dataset of http://lib.stat.cmu.edu/DASL/Datafiles/ceodat.html. On the *right*, a histogram of the medians of 10,000 bootstrap replicates of this data. This simulates the effect of sampling variation on the median; see Worked example 6.5

Worked example 6.5 (The Bootstrap Standard Error of the Median) You can find a dataset giving the salaries of CEO's at small firms in 1993 at http://lib.stat.cmu.edu/DASL/Datafiles/ceodat.html. Construct a 90% confidence interval for the median salary.

Solution Salaries are in thousands of dollars, and one salary isn't given (we omit this value in what follows). Figure 6.3 shows a histogram of the salaries; notice there are some values that look like outliers. This justifies using a median. The median of the dataset is 350 (i.e. $350,000—this is 1993 data!). I constructed 10,000 bootstrap replicates. Figure 6.3 shows a histogram of the medians of the replicates. I used the matlab `prctile` function to extract the 5% and 95% percentiles of these medians, yielding the interval between 298 and 390. This means that we can expect that, for 90% of samples of CEO salaries for small companies, the median salary will be in the given range.

Procedure 6.3 (The Bootstrap) Estimate the standard error for a statistic S evaluated on a dataset of N items $\{x\}$.

1. Compute r bootstrap replicates of the dataset. Write the i'th replicate $\{x\}_i$ Obtain each by:
   - **(a)** Building a uniform probability distribution on the numbers $1 \ldots N$.
   - **(b)** Drawing N independent samples from this distribution. Write $s(i)$ for the i'th such sample.
   - **(c)** Building a new dataset $\left\{x_{s(1)}, \ldots, x_{s(N)}\right\}$.
2. For each replicate, compute $S(\{x\}_i)$.
3. Compute
$$\overline{S} = \frac{\sum_i S(\{x\}_i)}{r}$$
4. The standard error estimate for S is given by:
$$\text{stderr}\left(\{S\}\right) = \sqrt{\frac{\sum_i \left[S(\{x\}_i) - \overline{S}\right]^2}{r-1}}$$

6.3 You Should

6.3.1 Remember These Definitions

6.3.2 Remember These Terms

6.3.3 Remember These Facts

6.3.4 Use These Procedures

6.3.5 Be Able to

- Compute the standard error of a sample mean.
- Plot and interpret error bars.
- Compute a confidence interval for a population mean using a sample.
- Compute a confidence interval for a population median using bootstrap samples.

Problems

Estimating the Population Standard Deviation

6.1 We have a population $\{X\}$, and we study random samples of N items (drawn with replacement). We write any particular sample $\{x\}$. Now consider $\mathsf{std}\left(\{x\}\right)^2$. This is a random variable (because different random samples of data would produce different values, at random).

(a) Show that $\mathbb{E}\left[\mathsf{std}\left(\{x\}\right)^2\right]$ is equal to

$$\mathbb{E}\left[\sum_i (x_i - \mathsf{popmean}\left(\{X\}\right))^2/N\right] -$$

$$(2/N)\mathbb{E}\left[(\mathsf{mean}\left(\{x\}\right) - \mathsf{popmean}\left(\{X\}\right)) \sum_i (x_i - \mathsf{popmean}\left(\{X\}\right))\right] +$$

$$\mathbb{E}\left[(\mathsf{mean}\left(\{x\}\right) - \mathsf{popmean}\left(\{X\}\right))^2\right].$$

(here the expectation is over sampling, though this has nothing to do with the point).

(b) Now show that for any sample

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

(c) Now use the methods of Sect. 6.1.1 to show that

$$\mathbb{E}\left[(\mathsf{mean}\left(\{x\}\right) - \mathsf{popmean}\left(\{X\}\right))^2\right] = \frac{\mathsf{popsd}\left(\{X\}\right)^2}{N}.$$

(d) Now show that

$$\mathbb{E}\left[\mathsf{std}\left(\{X\}\right)^2\right] = \mathsf{popsd}\left(\{X\}\right)^2 \left(\frac{N-1}{N}\right).$$

Samples and Populations

6.2 The Average Mouse: You wish to estimate the average weight of a mouse. You obtain 10 mice, sampled uniformly at random and with replacement from the mouse population. Their weights are $21, 23, 27, 19, 17, 18, 20, 15, 17, 22$ grams respectively.

(a) What is the best estimate for the average weight of a mouse, from this data?
(b) What is the standard error of this estimate?
(c) How many mice would you need to reduce the standard error to 0.1?

6.3 Sample Variance and Standard Error: You encounter a deck of Martian playing cards. There are 87 cards in the deck. You cannot read Martian, and so the meaning of the cards is mysterious. However, you notice that some cards are blue, and others are yellow.

(a) You shuffle the deck, and draw one card. You repeat this exercise 10 times, replacing the card you drew each time before shuffling. You see 7 yellow and 3 blue cards in the deck. As you know, the maximum likelihood estimate of the fraction of blue cards in the deck is 0.3. What is the standard error of this estimate?
(b) How many times would you need to repeat the exercise to reduce the standard error to 0.05?

Confidence Intervals for Population Means

6.4 The Weight of Rats You wish to estimate the average weight of a pet rat. You obtain 40 rats (easily and cheaply done; keep them, because they make excellent pets), sampled uniformly at random and with replacement from the pet rat population. The mean weight is 340 grams, with a standard deviation of 75 grams.

(a) Give a 68% confidence interval for the weight of a pet rat, from this data.
(b) Give a 99% confidence interval for the weight of a pet rat, from this data.

6.5 The Weight of Mice You wish to estimate the average weight of a mouse. You obtain 10 mice, sampled uniformly at random and with replacement from the mouse population. Their weights are $21, 23, 27, 19, 17, 18, 20, 15, 17, 22$ grams respectively. Notice there are too few mice to use a normal model.

(a) Give an 80% confidence interval for the weight of a mouse, from this data.
(b) Give a 95% confidence interval for the weight of a mouse, from this data.

6.6 The Probability of a Female Birth In Carcelle-le-Grignon at the end of the eighteenth century, there were 2009 births. There were 983 boys and 1026 girls. You can regard this as a fair random sample (with replacement, though try not to think too hard about what that means) of births. If you map each female birth to 1 and each male birth to 0, the probability of a female birth is the population mean of this random variable. You have a sample of 2009 births.

(a) Using the reasoning and data above, construct a 99% confidence interval for the probability of a female birth.
(b) Using the reasoning and data above, construct a 99% confidence interval for the probability of a male birth.
(c) Do these intervals overlap? what does this suggest?

6.7 Carcinomas vs Adipose Tissue The UC Irvine Machine Learning data repository hosts a dataset giving various electromagnetic measurements for different kinds of breast tissue. You can find the data at http://archive.ics.uci.edu/ml/datasets/Breast+Tissue. It was submitted by JP. Marques de Sá and J. Jossinet.

(a) Using this data, construct a 99% confidence interval for the mean value of the I0 variable for tissue from a carcinoma.
(b) Using this data, construct a 99% confidence interval for the mean value of the I0 variable for adipose tissue.
(c) Do these intervals overlap? what does this suggest?

6.8 Wine The UC Irvine Machine Learning data repository hosts a dataset giving various measurements of wine from three different regions of Italy. You can find the data at http://archive.ics.uci.edu/ml/datasets/Wine. This data was submitted by S. Aeberhard and seems to have originally been owned by M. Forina

(a) Using this data, construct a 99% confidence interval for the mean value of the flavanoids variable for wine from region 1.
(b) Using this data, construct a 99% confidence interval for the mean value of the flavanoids variable for wine from region 3.
(c) Do these intervals overlap? what does this suggest?

Programming Exercises

6.9 Investigating the construction of confidence intervals The UC Irvine Machine Learning data repository hosts a dataset giving various measurements of abalone at https://archive.ics.uci.edu/ml/datasets/Abalone. This data comes from an original study by W.J. Nash, T.L. Sellers, S.R. Talbot, A.J. Cawthorn and W. B. Ford, called "The Population Biology of Abalone (*Haliotis* species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the North Coast and Islands of Bass Strait", Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288) (1994). The data was donated by S. Waugh. There are 4177 records.

We will use the Length measurement. We will assume that the 4177 records is the entire population. Compute the population mean.

(a) Draw 10,000 samples of 20 records at random with replacement. Use each sample to compute a centered 90% confidence interval for the population mean, using the t-distribution. For what fraction of samples does the true population mean lie inside the interval?
(b) Draw 10,000 samples of 10 records at random with replacement. Use each sample to compute a centered 90% confidence interval for the population mean, using the t-distribution. For what fraction of samples does the true population mean lie inside the interval?
(c) Draw 10,000 samples of 10 records at random with replacement. Use each sample to compute a centered 90% confidence interval for the population mean, using a normal model (which you really shouldn't, because the sample is too small). For what fraction of samples does the true population mean lie inside the interval?
(d) Now repeat the last two subexercises, but using only three records. What conclusion do you draw?

6.10 Investigating the construction of bootstrap confidence intervals The UC Irvine Machine Learning data repository hosts a dataset giving various measurements of abalone at https://archive.ics.uci.edu/ml/datasets/Abalone. This data comes from an original study by W.J. Nash, T.L. Sellers, S.R. Talbot, A.J. Cawthorn and W. B. Ford, called "The Population Biology of Abalone (*Haliotis* species) in Tasmania. I. Blacklip Abalone (*H. rubra*) from the North Coast and Islands of Bass Strait", Sea Fisheries Division, Technical Report No. 48 (ISSN 1034-3288) (1994). The data was donated by S. Waugh. There are 4177 records. We will use the Length measurement. We will assume that the 4177 records is the entire population. Compute the population median.

(a) Draw 10,000 samples of 100 records at random with replacement. Use each sample to produce a bootstrap estimate of a centered 90% confidence interval for the population median. For what fraction of samples does the true population median lie inside the interval?
(b) Draw 10,000 samples of 30 records at random with replacement. Use each sample to produce a bootstrap estimate of a centered 90% confidence interval for the population median. For what fraction of samples does the true population median lie inside the interval?
(c) Draw 10,000 samples of 10 records at random with replacement. Use each sample to produce a bootstrap estimate of a centered 90% confidence interval for the population median. For what fraction of samples does the true population median lie inside the interval?
(d) What conclusion do you draw?