Statistical Thinking for the 21st Century

Chapter 7

Sampling

One of the foundational ideas in statistics is that we can make inferences about an entire population based on a relatively small sample of individuals from that population. In this chapter we will introduce the concept of statistical sampling and discuss why it works.

Anyone living in the United States will be familiar with the concept of sampling from the political polls that have become a central part of our electoral process. In some cases, these polls can be incredibly accurate at predicting the outcomes of elections. The best known example comes from the 2008 and 2012 US Presidential elections, when the pollster Nate Silver correctly predicted electoral outcomes for 49/50 states in 2008 and for all 50 states in 2012. Silver did this by combining data from 21 different polls, which vary in the degree to which they tend to lean towards either the Republican or Democratic side. Each of these polls included data from about 1000 likely voters – meaning that Silver was able to almost perfectly predict the pattern of votes of more than 125 million voters using data from only 21,000 people, along with other knowledge (such as how those states have voted in the past).

7.1 How do we sample?

Our goal in sampling is to determine the value of a statistic for an entire population of interest, using just a small subset of the population. We do

this primarily to save time and effort – why go to the trouble of measuring every individual in the population when just a small sample is sufficient to accurately estimate the variable of interest?

In the election example, the population is all registered voters, and the sample is the set of 1000 individuals selected by the polling organization. The way in which we select the sample is critical to ensuring that the sample is *representative* of the entire population, which is a main goal of statistical sampling.

It's easy to imagine a non-representative sample; if a pollster only polled individuals whose names they had received from the local Democratic party, then it would be unlikely that the results of the poll would be representative of the population as a whole.

In general, we would define a representative poll as being one in which every member of the population has an equal chance of being selected. When this fails, then we have to worry about whether the statistic that we compute on the sample is *biased* - that is, whether its value is systematically different from the population value (which we refer to as a *parameter*). Keep in mind that we generally don't know this population parameter, because if we did then we wouldn't need to sample! But we will use examples where we have access to the entire population, in order to explain some of the key ideas.

It's important to also distinguish between two different ways of sampling: with replacement versus without replacement. In sampling *with replacement*, after a member of the population has been sampled, they are put back into the pool so that they can potentially be sampled again. In *sampling without replacement*, once a member has been sampled they are not eligible to be sampled again. It's most common to use sampling without replacement.

7.2 Sampling error

Regardless of how representative our sample is, it's likely that the statistic that we compute from the sample is going to differ at least slightly from the population parameter. We refer to this as *sampling error*. The value of our statistical estimate will also vary from sample to sample; we refer to this distribution of our statistic across samples as the *sampling distribution*.

Table 7.1: Example means and standard deviations for several samples of Height variable from NARPS

sampleMean	sampleSD
167	9.123
171	8.345
170	10.604
166	9.519
168	9.538

Sampling error is directly related to the quality of our measurement of the population. Clearly we want the estimates obtained from our sample to be as close as possible to the true value of the population parameter. However, even if our statistic is unbiased (that is, in the long run we expect it to have the same value as the population parameter), the value for any particular estimate will differ from the population estimate, and those differences will be greater when the sampling error is greater. Thus, reducing sampling error is an important step towards better measurement.

We will use the NHANES dataset as an example; we are going to assume that the NHANES dataset is the entire population, and then we will draw random samples from this population. .

In this example, we know the adult population mean (168.353) and standard deviation (10.162) for height because we are assuming that the NHANES dataset *is* the population. Now let's take a few samples of 50 individuals from the NHANES population, and look at the resulting statistics.

As shown in Table 7.1, the means and standard deviations of our five random samples are similar. But they are not exactly equal to the population mean and standard deviation.

Now let's take a larger number of samples of 50 individuals, compute the mean for each sample, and look at the resulting sampling distribution of means. We have to decide how many samples to take in order to do a good job of estimating the sampling distribution – in this case, let's take 5000 samples so that we are really confident in the answer. The histogram in Figure 7.1 shows that the means estimated for each of the 5000 samples of 50 individuals vary

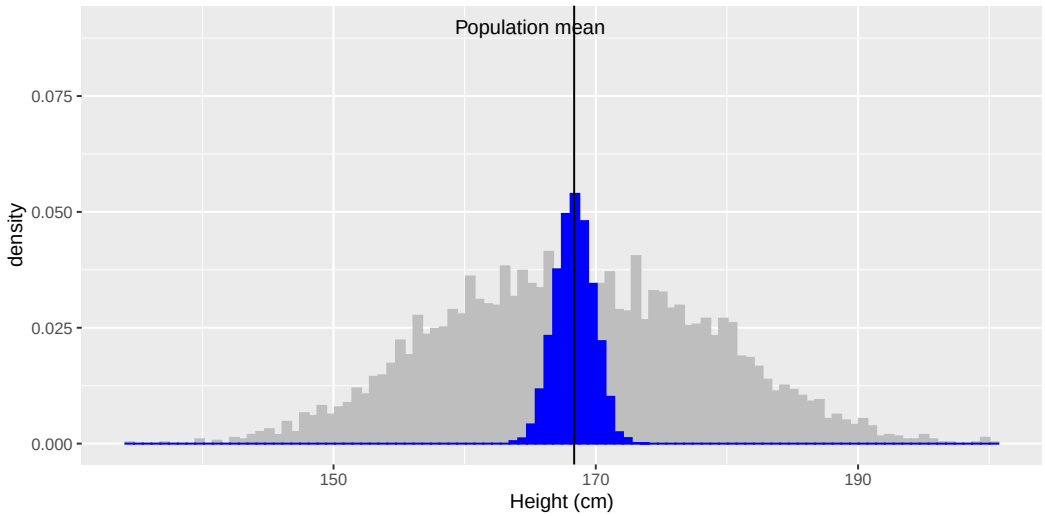

Figure 7.1: The blue histogram shows the sampling distribution of the mean over 5000 random samples from the NHANES dataset. The histogram for the full dataset is shown in gray for reference.

somewhat. But overall the samples are centered around the population mean. The average of the 5000 sample means (168.35) is very close to the true population mean (168.35).

7.3 Standard error of the mean

Often, we want to characterize how variable our samples are so that we can make inferences about the sample statistics. For the mean, we do this using a quantity called the *standard error* of the mean (SEM), which one can think of as the standard deviation of the sampling distribution. To compute the standard error of the mean for our sample, we divide the estimated standard deviation by the square root of the sample size:

$$SEM = \frac{\hat{\sigma}}{\sqrt{n}}$$

Note that we have to be careful about computing SEM using the estimated standard deviation if our sample is small (less than about 30).

The formula for the standard error of the mean says that the quality of our measurement involves two quantities: the population variability (our estimated standard deviation), and the size of our sample.

Because the sample size is the denominator in the formula for SEM, a larger sample size will yield a smaller SEM when holding the population variability constant. We have no control over the population variability, but we *do* have control over the sample size. Thus, if we wish to improve our sample statistics (by reducing their sampling variability) then we should use larger samples.

However, the formula also tells us something very fundamental about statistical sampling – namely, that the utility of larger samples diminishes with the square root of the sample size.