Chapter 6

Sampling

In this chapter and the next, we finally begin working with statistics in the way that is most useful for testing hypotheses. Unfortunately, *The Cartoon Guide to Statistics* gets a little redundant and perhaps too abstract in these two chapters. For our purposes, it is all a lot simpler than the presentation in the text, so you should strongly consider following the hints in this study guide so you can know what you need to know without getting lost in the formulas of this chapter.

6.1 Vocabulary

The vocabulary list for this chapter is short because it uses much of the vocabulary from previous chapters. Be sure you understand the following terms, but also be sure you can relate the vocabulary of previous chapters to this one.

- simple random sample
- standard error
- estimator
- central limit theorem
- t-distribution
- 95% confidence interval

6.2 Equations, specific rules, etc.

In the previous chapter, we found how easy it is to calculate properties of binomial and normal distributions. In this chapter and the next, we use these distributions and their easily calculated properties. The rules are simple; you just have to know when to apply them!

- For a binomial distribution where the true probability of success in the population is p_0 , the probability of success found in the sample is an estimate of the probability of success in the full population

$$\hat{p} \approx p_0 = E(\hat{P})$$

where $\hat{P}$ is the random variable representing the probability of success found in each of multiple samples of the same population. The distribution of $\hat{P}$ is approximately normal for large sample sizes and has the standard deviation

$$\sigma_{\hat{P}} = \frac{\sqrt{p_0(1-p_0)}}{\sqrt{n}}$$

- For a normal distribution of mean μ_0 and standard deviation σ_0, the mean of the sample is an estimate of the mean of the full population

$$\bar{x} \approx \mu_0 = E(\bar{X})$$

where $\bar{X}$ is the random variable representing the mean found in each of multiple samples of the same population. The distribution of $\bar{X}$ is approximately normal (the Central Limit Theorem) and has the standard deviation

$$\sigma_{\bar{X}} = \frac{\sigma_0}{\sqrt{n}}$$

That's it! The two ideas are very simple, but there's a lot behind them! Be sure to read the 'where' clauses above as they provide the explanation for why we can use these two estimators so much.

6.3 Specific notes

- p. 89: This page is important. It marks a transition. Make sure you understand that all we have done so far is just a foundation for this and future chapters.

- pp. 90-91: These pages are just trying to give you the motivation for the chapter. There is no real content here to worry about.

- pp. 92-93 Sampling design is very important! Almost every statistical tool we learn about from this point on depends on having chosen a simple random sample. There is nothing difficult in the concept of a simple random sample. However, it can be difficult to actually get a truly random sample. You should always keep in mind the dependence of statistical analyses on random sampling.

- pp. 94-97: These pages give several ways of attempting to create random samples. Again, pay close attention to the two words of warning on pp. 96-97.

- pp 98-101: Now we get past the warnings to the statistics. The text presents this in two sections, corresponding to the two items in the "Equations, specific rules, etc." section above. First is the situation where you are trying to estimate the relative frequency of some characteristic in the full population by taking a sample and finding the relative frequency of that characteristic in the sample. This is an obvious and simple idea.

  The point of all the rambling in the text is to give you some justification for actually believing that the relative frequency you find in the sample *is* a good estimate of the relative frequency in the full population.

  If you don't care about the justification, skip to p. 103.

  If you do care, here it is: Your sample is just one of many that could be taken. If you were to take lots and lots of samples, you would find an approximately normally distributed set of relative frequencies $\hat{P}$. The mean of this $\hat{P}$ distribution approaches the actual relative frequency in the full population. The standard deviation of this $\hat{P}$ distribution is just the standard deviation of a binomial distribution divided by n (to make it a relative frequency). This yields the formula given above.

- p. 102: This summary of the example doesn't really lead to any insights except to emphasize that it is the sample size which determines the spread of the estimator and therefore the error of your estimate. Increasing sample size reduces error.

- p. 103: The four-step process here really hides an important problem which the text addresses in chapter 7. The problem is that you don't know p, so you can't use it to come up with a "sampling error" which they denote as $\sigma(\hat{p})$. It is better not to talk about "sampling error" at all and just stick to standard error as it is described in chapter 7.

- pp. 104-105: Here is the second section. In this case, you are trying to estimate the mean value of some distribution. Again, the rambling of the text is meant to provide a justification for believing that you can estimate the mean value by just looking at a sample of the population and using the mean of that sample.

  If you don't care about the justification, skip to p. 106.

  If you do care, here it is: Your sample is just one of many that could be taken (sound familiar?). If you were to take lots and lots of samples, you would find an approximately normally distributed set of mean values $\bar{X}$. The mean of this $\bar{X}$ distribution approaches the mean of the full population. The standard deviation of this $\bar{X}$ distribution can be expressed in terms of the standard deviation of the full population as listed in the "Equations, specific rules, etc." section above, but again this is not very useful since we do not know the standard deviation of the full population. Again, this problem will be addressed in chapter 7. The main point is to notice that, once again, the spread of the estimator is a function of sample size. Increasing sample size reduces error.

- p. 106: The central limit theorem is an important result with which you should just play around until you are convinced of its truth.

- pp. 107-109: These pages are just a preview of chapter 7. Skim them, but don't dwell on them. It will all be more clear in chapter 7 (or at least in the study guide to chapter 7).

6.4 Exercises

Exercise 6.1: Assume that it is known that the proportion of left-handed people in a certain population is 25%. How many left-handed people would you expect to find in a sample of 100 people from this population? If you took many different samples, you would get different numbers of left-handed people in each sample. What is the smallest number you would expect (at a level of 95% confidence)? Hint: you'll need a standard deviation for the whole population, from which you can calculate the standard deviation of the distribution of samples.

Exercise 6.2: Assume that it is known that the height of women in a population is described by a normal distribution with a mean of 68 inches and a standard deviation of 4 inches. What are the expected values for the mean and standard deviation of a sample of 100 women from this population? What are the expected values for the mean and standard deviation of a sample of 10,000 women from this population?

Note: The exercises in this chapter have to start with "Assume that it is known...." What we really want to be able to do is get rid of this assumption and say something about unknown characteristics of the full population by using the sample. That is exactly what we will do in chapter 7!