1. Overview

In previous chapters, we examined the variability associated with the sum of a box (Chapter 17) and of percentages (Chapters 20 and 21). In chapter 23, we turn to the variability of sample averages.

2. Example

Suppose we draw a simple random sample of size n from a large population. An example might be -- draw a simple random sample (SRS) of 100 American women from the population of all American women. Measure their heights. Suppose we knew that the population has mean of 5'3" and a standard deviation of 2.5". (suppose you know what the box looks like for now)

Then each woman drawn from the "box" has an expected height of 5'3" with a SE = 2.5"; in other words, each woman is expected to be like the original distribution. For the sample of 100, the expected value for the average of 100 draws is simply equal to the average of the box. You could also think of it as the sum of the draws divided by the number of draws, but again, that's just the average. (see page 410)

The standard error of the average (not sum or percentage) for the sample of 100 is

$$\frac{\sqrt{\text{number of draws}} * SD}{\text{number of draws}}$$ the number of draws is the same as sample size.

for a sample of 100 from this particular box the standard error would be:

$$\frac{\sqrt{100} * 2.5}{100}$$ = .25 inches notice we used the population SD for the box SD.

3. Interpretations

When you draw just one woman at random from the "box", your best guess about her height is 5'3" and there is a 68% chance that she will be within 2.5" of that value, a 95% chance that she will be within 5" of that value.

But when you are drawing 100 women from the box at random. Your best guess about their average height is still 5'3", but now there is a 68% chance that you will be within .25 inches of the population average. And there is a 95% chance of being within 1/2 inches of the population average.

Why? The average of an entire sample has a much better chance of being close to the parameter than just a sample of size 1. You already "know" in your "gut" that having a larger sample size is preferred if you are trying to estimate a percentage or an average.

We can make these probability statements for the average of draws from a box even when the underlying population is not normally distributed. It's the average of all of the samples (in theory) that are normally distributed. This works when your samples are reasonably large (30 or more is reasonable)

4. Properties

The expected value of the sample average is the population average. (see page 410)

Looking at the standard error of the sample average again

$$\frac{\sqrt{\text{number of draws}} * SD}{\text{number of draws}}$$ where the number of draws is the same as sample size, (also page 410)

Thus, the standard error of a sample, say with twenty people, will be smaller than the standard deviation for individual measurements. It's easier to predict the average for a group than it is to predict a single measurement.

The probability histogram, that is, the distribution of sample means, will follow a normal curve even if the population it comes from does not. But for this to be true, the samples drawn must be of a reasonable size (> 30 at least).

5. When the population mean and standard deviation are unknown (23.2)

This is like the material presented in Chapter 21 and reflects real life. Usually, you don't know parameters and can't really measure them. But you may have a good sample. Just like chapter 21, you use sample information to make statements about the population. Once again, constructed in the form of confidence intervals like Chapter 21.

Suppose a psychologist wants to know the average IQ of the 28,000 students at USC. Suppose he takes a simple random sample of 100 and the sample average turns out to be 95. The standard deviation of the sample is 50.

The average IQ of all USC students is estimated as 95, but of course there is always chance error when you are dealing with samples. He will want to put a ± estimate around the 95.

To do that, he will need an SE. Things to do

1. Find the SE for the sum of draws:

   $\sqrt{100} * 50 = 500$, notice I used the sample standard deviation as an estimate of the population (or box) SD

2. Find the SE for the average:

$$\frac{\sqrt{\text{number of draws}} * SD}{\text{number of draws}} = \frac{\sqrt{100} * 50}{100} = 5$$

3. Construct the confidence interval, it has the form:

$$\bar{x} \pm Z * (SE_{average})$$

95 ± (1*5) (for 68% confidence, which is one standard error)
95 ± (2*5) (or ± 10 for 95% confidence, which is 2 standard errors -- what would 99% confidence look like?)

In about 68% of all samples of size 100, if you go ± 5 IQ points from the sample average of 95, you will cover the USC population average. In about 95% of all samples of size 100, if you go ± 10 IQ points from the sample average of 95, you will cover the USC population average. Or, you might make statements of confidence: "I am 68% confident that the range 90 to 100 covers the true USC IQ average" or "I am 95% confident that the range 85 to 105 covers the true USC IQ average". Human IQ average 100, so it is entirely possible that the true mean IQ of USC students is 100.

Remember that the normal curve is a good approximation of the distribution of sample averages if you could sample again and again. It allows you to make probability statements.

6. Things to keep in mind

A. Am I moving forward from a known box? If yes, I can probably make some strong statement about a sample. (statement of chance or probability) this is Chapter 17 & 20 and part of 23.
B. If the box is unknown, I'm moving backward from a sample and cannot make as strong of a statement about the parameter. (confidence interval) this is Chapter 21 and part of 23.
C. When you don't know much about the original population, the distribution of sample averages will be normal, but the underlying original population is not necessarily normal. (Chapter 23)

I. What type of sample statistic are you being asked about?

	Count or Sum or Total	Proportion or Percentage (which is a proportion*100)	Mean or Average
Expected Value	Number of draws * box average (page 289)	Box percentage (page 359)	Box Average (page 410)
Standard Error	$\sqrt{draws} * SD_{box}$ (page 291)	$\frac{\sqrt{draws} * \sqrt{fractionof\,1's * fractionof\,0's}}{draws} * 100$ see page 360	$\frac{\sqrt{draws} * SD_{box}}{draws}$ (page 410)
Notes	The box could be a one-zero box, but generally it's a box that contains different kinds of numbers (see Chapter 17). Assumes sampling w/ replacement.	For one-zero boxes only. Assumes sampling w/ replacement.	Generally the average and SD are given and do not need to be calculated. Assumes sampling w/ replacement.

II. How are you being asked to apply this statistic?

	Count or Sum or Total	Proportion or Percentage (which is a proportion*100)	Mean or Average
Using the normal curve (review Chapter 5)	Find a Z score then the area from Table A-105 using: $Z = \frac{observed - expected}{SE_{sum}}$ (page 294-296)	Find a Z score then the area from Table A 105 using: $Z = \frac{observedpercentage - expectedpercentage}{SE_{percentage}}$ (page 362-366)	Find a Z score then the area from Table A105 using: (page 410-411) $Z = \frac{observedmean - expectedmean}{SE_{mean}}$
Calculating confidence intervals	Not done in this textbook	Page 381: $\begin{matrix} sample \\ percentage \end{matrix} \pm multiplier * SE_{percentage}$	Page 416-417 $\begin{matrix} sample \\ average \end{matrix} \pm multiplier * SE_{average}$
Hypothesis Testing - use Z test	See Chapter 26.5	See Chapter 26.5	$Ztest = \frac{observedmean - hypotheticalmean}{SE_{mean}}$ use this to find the area from Table A105, area values as extreme or more extreme than the Z result are called "p-values" (page 482) p variables smaller than 5% are considered statistically significant and lead us to reject the null hypothesis (page 484)