Math 140
Introductory Statistics

Professor Bernardo Ábrego
Lecture 28
Sections 8.4, 8.5, 9.1

Example E65 (page 537)

- A poll of 256 boys and 257 girls age 12 to 17 asked, "Do you feel like you are personally making a positive difference in your community?" More girls (195) than boys (161) answered "yes."
  - a. Using a one-sided test, is this a statistically significant difference? That is, if all teens were asked, are you confident that a larger proportion of girls than boys would say "yes"? Assume that the samples were selected randomly.
  - b. The report says, "Participants were selected through random digit dialing." Do you have any concerns about whether such a procedure would give a random sample?
  - c. Find a 95% confidence interval for the proportion of all teens who would answer yes. What additional assumption do you need to make to do this?

9.1 A Confidence Interval for a Mean

- It will have the same general form as the one for proportions.

$$\hat{p} \pm z^* \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

- In other words

$$statistic \pm (critical\ value) \cdot (standard\ deviation\ of\ statistic)$$

Example: Average Body Temperature

- To determine an up-to-date average of body temperature, researchers took the body temperatures of 148 people at several different times during two consecutive days. A portion of these data, for ten randomly selected women, is given here (in °F):

97.8 98.0 98.2 98.2 98.2 98.6 98.8 98.8 99.2 99.4

- The mean body temperature, $\overline{x}$, for this sample of ten women is 98.52, and the standard deviation, s, is 0.527. Are these statistics likely to be equal to the mean μ and standard deviation σ for the population? How can you determine the plausible values of the mean temperature of all women?

Solution Try

- $\overline{x}$ *and* s *are not exactly equal to* the population parameters μ and σ.
- Plausible values of the mean body temperature of all women, μ, are those values that lie "close" to $\overline{x} = 98.52$, where "close" is defined in terms of standard error.
- The standard error of the sampling distribution of a sample mean (Section 7.3) is given by

$$SE_{\overline{x}} = \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the population and n is the sample size. When the sample size is large enough or the population is normally distributed, in 95% of all samples, $\overline{x}$ and μ are no farther apart than 1.96 times the standard error.

Solution Try (We don't know σ)

- The standard error of the sampling distribution of a sample mean (Section 7.3) is given by

$$SE_{\overline{x}} = \sigma_{\overline{x}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the population and n is the sample size. When the sample size is large enough or the population is normally distributed, in 95% of all samples, $\overline{x}$ and μ are no farther apart than 1.96 times the standard error.

- So plausible values of μ lie in the interval

$$\overline{x} \pm 1.96 \cdot \frac{\sigma}{\sqrt{n}} \text{ or } 98.52 \pm 1.96 \cdot \frac{??}{\sqrt{10}}$$

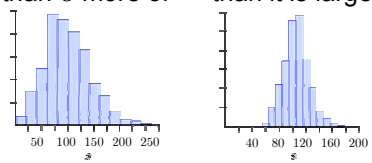
- **We don't know σ !!** (and we shouldn't expect to know it)
  What do we do?

The effect of estimating σ

- In real applications, you almost never know the true population standard deviation σ
- What can we do?
- **Answer**: We have to use the sample standard deviation s as an estimate.
- How will making that change—substituting s for σ—affect your inferences?
  - Some samples give an estimate that's too small: s < σ. Others give an estimate that's too big: s > σ.
  - On average the small and large values even out so that the sampling distribution of s has its center very near σ.

The effect of estimating σ

- Although s is about equal to σ on average, it tends to be smaller than σ more often than it is larger.



Display 9.1 Approximate sampling distribution of s for samples of size 4 (left) and size 20 (right) for a normally distributed population with $\sigma = 107$.

- This is because the sampling distribution of s is skewed right. The sampling distribution of s becomes less skewed and more approximately normal as the sample size increases.

How to adjust for estimating σ

- Estimating the standard deviation does not affect the center of a confidence interval (the center is at the sample mean $\bar{x}$).
- Substituting s for σ **does lower the overall capture rate** unless you compensate by increasing the interval widths by replacing z^* with a larger value, t^*.
- Questions: Which value t^*? How do we find it?

Student-Statistician Dialogue

- ***Student:*** Where does the value of t^* come from?
- ***Statistician:*** In principle, you could find it using simulation. Set up an approximately normal population, take a random sample, compute the mean and standard deviation. Do this thousands of times. Then use the results to figure out the value of t^* that gives a 95% capture rate for intervals of the form $\bar{x} \pm t^* \cdot s / \sqrt{n}$
- ***Student:*** Wouldn't that take a lot of work?
- ***Statistician:*** Yes, especially if you went about it by trial and error. Fortunately, this work has already been done, long ago. A statistician, W. S. Gosset (English, 1876–1937), who worked for the Guinness Brewery, actually did this back in 1915. Four years later, the geneticist and statistician R. A. Fisher (English, 1890–1962) figured out how to find values of t^* using probability theory. It turns out that the value of t^* depends on just two things—how many observations you have and the capture rate you want.

Student-Statistician Dialogue

- ***Student:*** So t^* doesn't depend on the unknown mean or unknown standard deviation?
- ***Statistician:*** No it doesn't, which is very handy because in practice you don't know these numbers. Suppose, for example, you have a sample of size n = 5 and you want a 95% interval. Then you can use t^* = 2.776 no matter what the values of μ and σ are.
- ***Student:*** Where did you get that value for t^*?
- ***Statistician:*** From a t-table, although I could have gotten it from a computer. A brief version of the table is shown in Display 9.6. Table B in the Appendix is more complete. The confidence level tells you which column to look in. For example, for a 95% interval, you want a tail area of .025 (half of .05) on either side, so you look in the column headed .025. For the row, you need to know the degrees of freedom, or *df* for short.

Student-Statistician Dialogue

- ***Student:*** Degrees of freedom? What's that?
- ***Statistician:*** There's a short answer, a longer answer, and a very long answer. The longer answer will come in E40. The very long answer is for another course. For the moment, here's the short answer: The **degrees of freedom** is the number you use for the denominator when you calculate the sample standard deviation. So for these confidence intervals, $df = n - 1$, where n is your sample size.
- If n = 5, for example, then df = 4 and you look in that row. If you turn to Table B in the Appendix and look in the row with df = 4 and the column with tail probability 0.025, you'll find the value 2.776 for t^*.

(Better than) Calculator Note

- To get t^* on your calculator you can use `TInterval` and the following values:
  - $\overline{x}$:0
  - Sx: $\sqrt{n}$
  - n: n
  - C-Level: Confidence Level