MATH 2510-001 Week 10

This week's material will be the most technical in the entire course. We need to finish building the conceptual (and mathematical) foundations for *statistical inference*. From prior experience, I know it's going to be tough.

> *"All of my life been wadin' in \\ Water so deep now we got to swim \\ Wonder will it ever end. \\ We're all together in the same boat \\ I know you, you know me \\ Baby, you know me \\ I just wanna dream." (Get Free, Major Lazer ft. Amber Coffman)*

Many sections are going to "wave their hands", flail, and end up cheating themselves of understanding this stuff. We will not, because, I repeat, this material constitutes the *foundations for statistical inference*, and *understanding inference* is arguably the *primary reason* to take this course. Here is a study guide with ratings of difficulty to help us get through the material safely:

- (□) should require the usual level of effort. No fancy footwork required.
- (♦) may require additional effort; considering slowing down and doing the suggested readings.
- (♦♦) may seem *spooky*. Here's a tip: approach this material *carefully* and ask for help. E.g., consider visiting the MARC (MATH 175) while reading section 7.4 in Brase and Brase.

Study Guide: Point Estimates and Sampling Variability

1. (□) Define sample statistic as a point estimate for a population parameter, for example, the sample proportion is used to estimate the population proportion, and note that point estimate and sample statistic are synonymous.
2. (□) Recognize that point estimates (such as the sample proportion) will vary from one sample to another, and define this variability as sampling variation.
3. (□) Calculate the sampling variability of the proportion, the standard error, as $SE = \sqrt{\frac{p(1-p)}{n}}$, where p is the population proportion.
    - Note that when the population proportion p is not known (almost always), this can be estimated using the sample proportion, $SE = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
4. (♦) Standard error (SE) measures the variability in point estimates from different samples of the same size and from the same population; i.e., SE measures the sampling variability.
5. (□) Recognize that when the sample size increases we would expect the sampling variability to decrease.
    - Conceptually: Imagine taking many samples from the population. When sample sizes are large the sample proportion will be much more consistent across samples than when the sample sizes are small.
    - Mathematically: $SE = \sigma/\sqrt{n} \approx s/\sqrt{n}$, when n increases, SE will decrease since n is in the denominator.

6. (⧫) Notice that sampling distributions of point estimates coming from samples that don't meet the required conditions for the Central Limit Theory (about sample size and independence) will not be normal.

* *Reading: Section 5.1 of this packet*
* *Reading: Section 7.3 of Understandable Statistics (on WebAssign)*
* *Test yourself:*

    1. *For each of the following situations, state whether the variable is categorical or numerical, and whether the parameter of interest is a mean or a proportion: (a) In a survey, college students are asked whether they agree with their parents' political ideology. (b) In a survey, college students are asked what percentage of their non-class time they spend studying.*
    2. *Explain what is going on in Figures 5.4 and 5.5 of this packet (pages 176 and 177).*

Study Guide: Confidence Intervals for a Proportion

7. (⧫) Define a confidence interval as the plausible range of values for a population parameter.
8. (□) Define the confidence level as the percentage of random samples which yield confidence intervals that capture the true population parameter.
9. (□) Calculate an approximate 95% confidence interval by adding and subtracting 2 standard errors to the point estimate: $\textit{point estimate} \pm 2 \times SE$.
10. (□) Recognize that the Central Limit Theorem (CLT) is about the distribution of point estimates, and that given certain conditions, this distribution will be nearly normal.

    - In the case of the proportion the CLT tells us that if

      (1) (⧫) the observations in the sample are independent, and

      (2) (□) there are at least 10 successes and 10 failures[1]

      then the distribution of the sample proportion will be nearly normal, centered at the true population proportion and with a standard error of $\sqrt{\frac{p(1-p)}{n}}$.

$$\hat{p} \sim N\left(\mu_{\hat{p}} = p, SE = \sqrt{\frac{p(1-p)}{n}}\right)$$

    - (□) When the population proportion p is unknown, condition (2) can be checked using the sample proportion $\hat{p}$.

11. Recall that independence of observations in a sample is provided by random sampling (in the case of observational studies) or random assignment (in the case of experiments).

    - In addition, the sample should not be *too* large compared to the population, or more precisely, should be smaller than 10% of the population, since samples that are too large will likely contain observations that are not independent.

[1] WebAssign only requires 5, ok? But 5 doesn't fly in social science, journalism, business, the natural sciences, etc. You need 10 successes and failures if you want to apply the CLT like a professional.

12. Recognize that the nearly normal distribution of the point estimate (as suggested by the CLT) implies that a more precise confidence interval can be calculated as

$$point\ estimate \pm z^\star \times SE,$$

where $z^\star$ corresponds to the cutoff points in the standard normal distribution to capture the middle XX% of the data, where XX% is the desired confidence level.

- For proportions this is $\bar{x} \pm Z^\star \sqrt{\frac{p(1-p)}{n}}$.
- Note that $z^\star$ is always positive.

13. Define margin of error as the distance required to travel in either direction away from the point estimate when constructing a confidence interval, i.e. $z^\star \times SE$.

- Notice that this corresponds to half the width of the confidence interval.

14. Interpret a confidence interval as "We are XX% confident that the true population parameter is in this interval", where XX% is the desired confidence level.

- Note that your interpretation must always be in context of the data – mention what the population is and what the parameter is (mean or proportion).

* *Reading: Section 5.2 of this packet*
* *Reading: Section 7.3 of Understandable Statistics (on WebAssign)*
* *Test yourself:*

1. *Explain, in plain English, what is going on in Figure 5.6 of this packet (page 182).*
2. *List the conditions necessary for the CLT to hold. Make sure to list alternative conditions for when we know the population distribution is normal vs. when we don't know what the population distribution is, and the when the sample size is barely over 30 vs. when it's very large.*
3. *Confirm that $z^\star$ for a 98% confidence level is 2.33. (Include a sketch of the normal curve in your response.)*
4. *Explain, in plain English, the difference between standard error and margin of error.*
5. *Suppose 10% of CU Boulder students smoke. You collect many random samples of 100 CU Boulder students at a time, and calculate a sample proportion ($\hat{p}$) for each sample, indicating the proportion of students in that sample who smoke. What would you expect the distribution of these $\hat{p}$'s to be? Describe its shape, center, and spread.*

Study Guide: Estimating Differences of Population Means and Proportions

15. For independent populations, the confidence intervals for the difference of two means or of two populations follow the same format as confidence intervals for a single mean or a single proportion.

16. The margin of error "depends on the sampling distribution"[2] of the point estimate:

(a) $\bar{x}_1 - \bar{x}_2$ for $\mu_1 - \mu_2$ (in the case we want a confidence interval for a difference of means)

- When σ_1 and σ_2 are known, the standard error is

$$SE = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

and the margin of error is $E = z_c \cdot SE$ (please know how to find the critical z-value z_c with `invNorm`).

- When σ_1 and σ_2 are unknown, and $n = \min\{n_1, n_2\}$ is relatively small (ie., less than 100), the standard error is found by guessing the sample standard deviation is a good enough estimator for the population standard deviation (e.g., guessing that $s_1 \approx \sigma_1$), then taking the standard error to be $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ and the margin of error to be $E = (t_n)_c \cdot SE$ with t_n found from Student's t distribution with $n - 1$ degrees of freedom, where $n = \min\{n_1, n_2\}$ is the minimum of the sample sizes. (Please know how to find the critical $(t_n)_c$-value with `invT`).
- When $n = \min\{n_1, n_2\}$ is relatively large (larger than 100), you have enough data to accurately "guess" σ_1 and σ_2. Refer to the first case.

(b) $\hat{p}_1 - \hat{p}_2$ for $p_1 - p_2$ (in the case we want a confidence interval for a difference of proportions)

- When n_1 and n_2 sufficiently large (e.g., when $n_1\hat{p}_1$ and $n_1(1-\hat{p}_1)$ are both larger than 10, and likewise, for the second proportion, $n_2\hat{p}_2$ and $n_2(1-\hat{p}_2)$ are *also* both larger than 10) here is the standard error of the distribution of the difference in two independent sample proportions[3]

$$SE_{(\hat{p}_1 - \hat{p}_2)} = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

and here is the margin of error for the corresponding confidence interval

$$E_{(\hat{p}_1 - \hat{p}_2)} = z_c \cdot SE_{(\hat{p}_1 - \hat{p}_2)}$$

- When n_1 and n_2 are small, you need a larger sample!

∗ *Reading: Section 6.2 of this packet*

∗ *Reading: Section 7.4 of Understandable Statistics (WebAssign)*

∗ *Test yourself:*

1. *Suppose a 95% confidence interval for the difference between the CU Boulder and CSU students who smoke (calculated using $\hat{p}_{CU} - \hat{p}_{CSU}$) is (-0.08,0.11). Interpret this interval, making sure to incorporate into your interpretation a comparative statement about the two schools.*
2. *Does the above interval suggest a significant difference between the true proportions of smokers at the two schools?*

[2] The "sampling distribution" here is *either* the distribution of a linear combination of sample means $\bar{X}$, $\bar{Y}$, etc. (when we are interested in the difference of sample means) *or* the distribution of a linear combination of point estimates $\hat{p}$, $\hat{\ell}$, etc. (when we are finding a sample proportion).

[3] This is the standard error also for hypothesis tests when $H_0 : p_1 - p_2 =$ a nonzero value.