Lab 5: Power and Sample Size

Javier E. Flores

March 11, 2019

Introduction

In our last lecture, we ended the discussion with a brief preview of the concept of **power**. As a reminder, we defined the power of a test as the probability of rejecting a false null hypothesis. Throughout this lab, we'll be exploring power and its properties. A large focus of this lab will also be on sample size. This is particularly important when considering the data collection process and its impact on statistical inference. There is often a balance that must be struck between the cost - either measured monetarily, by time invested, or some other way - of collecting data and the return those data provided in allowing you to draw inference. In today's lab, we'll learn how to strike this balance and find the minimum sample size necessary to achieve certain inferential goals.

As you work through the lab with your group, you will be asked to answer several questions. Please submit your responses (as a group) in a single, separate document. Include the original questions as well as your group's response in the final submission.

Interval Estimation

We'll begin this lab by first revisiting confidence intervals. Recall that, assuming a normal distribution, the P% confidence interval for a single mean is given by:

$$\bar{x} \pm z_{crit}\frac{\sigma}{\sqrt{n}},$$

where z_{crit} refers to the appropriate critical value on the standard normal distribution. Recall also that we call $z_{crit}\frac{\sigma}{\sqrt{n}}$ the interval **margin of error (MOE)**, i.e.

$$MOE = z_{crit}\frac{\sigma}{\sqrt{n}}$$

Q1) Using the information above, express n as a function of z_{crit}, MOE, and σ.

Recall that we use the t-distribution rather than the normal distribution in estimating σ when the sample size is small. We also learned that the t-distribution shape varies with the degrees of freedom.

Q2) Would the formula you derived in the previous question be valid if you swapped z_{crit} for t_{crit}? In other words, could the same general formula be used for t-distribution based confidence intervals AND z-distribution based confidence intervals? Why or why not?

Q3) Preliminary data on the regularly occurring non-stop flight from Boston to San Francisco, United 433, is provided below. The data describe the airtime in minutes for each of United 433's flights.

Airtime				
353	351	377	348	402
358	380	370	351	388
374	346	372	359	381
360	369	356	369	
346	374	407	384	
373	385	356	368	
363	377	360	398	

In 2016, United 433 was scheduled to depart at 6AM EST and arrive before 10AM PST. Due to the time zone difference, the expected flight time is 420 minutes. Using the preliminary data above, determine how large a sample needs to be collected to achieve a 95% confidence interval estimate for the mean airborne time that has a margin of error of 2 minutes. Assume that the data are normally distributed.

In addition to confidence intervals for the mean, we've also learned that an approximate P% confidence interval for a single proportion is given by:

$$\hat{p} \pm z_{crit}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Q4) Derive an expression for n as a function of $\hat{p}$, MOE, and z_{crit}.

Q5) Suppose you are interested in conducting a poll prior to the 2016 election in order to determine which presidential candidate, Hillary Clinton or Donald Trump, is preferred among Grinnellians. Our interest is in determining the proportion of Grinnellians that support Hillary Clinton. How large of a sample would be needed to estimate this proportion within 3% at the 95% confidence level? Assume that we have no preliminary information to inform what this proportion might be.

Q6) Assume that you found an earlier poll estimating the proportion of Hillary supporters to be 42%. How large a sample would be needed to achieve a 3% margin of error in this scenario?

Q7) Consider your answers to questions 5 and 6. Is there some other preliminary estimate that would lead to a larger sample size than what was found in question 5? If so, what is it? If not, why not?

Hypothesis Testing

We've learned that hypothesis testing is a useful framework for formally executing the scientific process. Given a set of data, we currently have the ability to measure the amount of evidence against some null hypothesis in order to draw a conclusion. However, we have yet to discuss what to do *before* data are collected and during the study design stage. During this stage, when you are planning your study, one of the most important considerations is how likely you are to reject the null hypothesis if it is truly false.

This is when knowing about power becomes important. As mentioned earlier, the power of a test is the probability of rejecting the null hypothesis when it is truly false. Assuming β represents the type II error rate, we can think of power as 1-β. If the power of a test is too low, we have little chance of obtaining a statistically significant result even if null hypothesis is false. In practice, we aim to achieve a power of at least 80%. The power of a test is dependent on three factors:

1) effect size, or how different the true parameter is from the null value

2) sample size

3) significance level, or type I error rate

Designing a study with at least 80% power requires that we first understand the role each of these factors has on power. We will use the following app in order to gain an understanding of power and how it relates to each of these three factors.

Q8) Opening the app, what does the blue distribution represent? What does the green distribution represent? What do the dashed vertical lines represent?

Q9) Why is the light blue region the probability of making a type I error? Why is the dark green region the probability of making a type II error?

Q10) Using the app, specify $\alpha = 0.05$, $\mu_1 = 0$, and $\sigma_1 = 1$. Create a table which records the power for each of the following effect sizes: 2.1, 2.5, 3, 3.5, 4, 4.5.

Q11) Using the app, specify $\alpha = 0.05$, $\mu_1 = 0$, $\sigma_1 = 1$, $\mu_2 = 3$, and $\sigma_2 = 1$. Create a table which records the power for each of the following significance levels: 0.01, 0.03, 0.05, 0.10, 0.15, 0.20.

Q12) With an increase in sample size, do you expect the power to increase or decrease? Why? (Hint: Think about how the standard error is related to the sample size!)

Q13) In no more than three sentences, summarize the effect that individually varying the effect size, sample size, and significance level has on the power of a test. In other words, for each factor, describe how the power change by varying the factor and keeping all others constant.

Challenge (Optional)

In R, write a function that produces a plot similar to what is shown in the power app. The arguments of this function should be the significance level, null distribution mean, null distribution standard deviation, alternative distribution mean, and alternative distribution standard deviation. The closer your output is to the app output, the more points will be awarded. A maximum of 6 points will be given for completing this challenge.

You may work on this challenge as a group, but I expect individuals to turn in separate copies of their own code and generated statistics/figures.

Turning in this challenge may prompt a brief assessment in which you will be asked to state the code used to generate any one of the results produced.

This challenge may be turned in and redeemed for credit at any point prior to the second exam, which will be held two weeks after we return from Spring Break