Section 1: Study Design + Hypothesis Testing

TF: Reagan Rose (rrose@g.harvard.edu)

September 8, 2015

Elements of Study Design

A key aspect of Stat 139 is ensuring that you understand the subtleties of designing a study and drawing appropriate inferences from a particular design. What is "study design"? It generally consists of the following:

1. **Data collection:** How is the data collected? What kind of data is it (experimental, observational)? What kind of assumptions can we make about the data (i.e. is it normally distributed?)?
2. **Sampling method:** Who do we sample, and how big of a sample is sufficient? Are some samples better than others?
3. **Assignment mechanism:** In an experiment, how do we decide which subjects receive *treatment* and which receive *control*?

Why does it matter? The study design determines how you can interpret your results, so understanding study design will help you make true sense of your data (and avoid making inappropriate inferences [1]

Data Collection

When analyzing data, it is important to understand how the data was collected. There are two primary methods of data collection to be familiar with:

1. **Observational Studies:** Data is collected based on what is *observed* in the natural environment. There is no ability to manipulate any variables or conditions. Most importantly, we do not have a control group.
   Examples:_____

[1] see the American Journal of Cardiology article on "Why Jogging Will Literally Kill You" for an example.

2. **Experiments:** Data is collected by assigning subjects to treatment and control groups and then observing the results.
   Examples:_____

Why does it matter? Because good experiments can tell us something about *causation*, while good observational studies can only tell us about *correlation*.

Types of Data

1. **Continuous:** The can be any number in a large range.
2. **Nominal:** The value is a category with no natural ordering (i.e. hair color, sex)
3. **Ordinal:** There is some natural order to the categories (i.e. letter grades, places in a race).

Why does it matter? Again, we just need to understand the types of data we have in order to conduct appropriate analyses.

Statistics, Estimates, Parameters

Statisticians are known to be unsually picky about semantics. Here's the run down:

1. **Parameter:** A characteristic of the population, or the *true value* we want to measure such as average height of the world population).
2. **Statistic:** A *function* of the data which tells you something about the data (i.e. mean, median, variance).
3. **Estimator:** A specific type of statistic that yields a value (hopefully) close to the population parameter (i.e. sample mean is an estimator for population mean).
4. **Estimate:** A *realization* of an estimator, or the actual value you get when you plug in your data to your estimator.

Why does it matter? Understanding these terms and their differences is essential to understanding the theory and practice of statistics.

Types of Sampling

:

1. **Simple Random Sample:** The most basic method of picking a sample is to select a sample size n (usually based on the *power* of the test you'd like to conduct) and pick units with equal probability.
2. **Systematic Random Sample:** A variant of simple random sampling, this introduces some pattern to the sampling scheme (i.e. pick every 5th person).

3. **Stratified Random Sample:** A generalization of random sampling in which we divide the population into strata and do SRS within each strata. This ensures that if there are distinct groups of people in our population, we sample them proportionally.

Why does it matter? In an ideal world, we would just run our experiments on the whole population. But this generally isn't possible (can you think of why?). Instead, we use sampling methods to determine how much data we actually need in order to say something valid about the whole population.

Hypothesis Testing

: Suppose you'd like to answer a question with respect to some scientific hypothesis, and you conduct an experiment to do so. How do we know whether to accept or reject the hypothesis?

1. Formulate the null hypothesis H_0 and alternative hypothesis H_a. The "null" hypothesis always corresponds to "no" (i.e. no difference between groups, no difference from normality, etc.).
2. Calculate a test statistic using your data (t-statistic, z-statistic, χ^2).
3. Compute the p-value of the test statistic based on an appropriate sampling distribution.
4. Compare the p-value to the desired significance level α. If $p < \alpha$, reject H_0 (if the p is low, reject the H_0).

Why does it matter? We want to be able to make claims about statistical significance. That is, we want to say if we believe something occured systematically or just due to chance. Hypothesis test give us a way to do that.

Exercises

Question 1. Suppose we want to examine whether GPA is related to one's love life for students in statistics. To that end, we spam our survey to every mailing list we can find and hope for responses.

(a) What is the target population? Who are our study units?

(b) What kind of variables should we be measuring, and how does this affect our analysis?

(c) List a couple of ways we could improve our study, and explain what statistical benefits they would serve.