MTH 2130 — Probability & Statistics

Final Project Guidelines

1 Directions

This final project is worth 15% of your course grade. Students will work in groups of three. Your write-up along with any plots, tables, and calculations, should be neatly and carefully written, ideally in LaTeXor some similar format. **Due: March 12th at noon**. Drop off hard copies of the project write-up to Paul Coveney's office in Milas Hall.

2 Elements to the Final Project

2.1 The "Question"

Hypothesis testing is one of the fundamental concepts we study in this course, so it is natural to begin your project with a question (or hypothesis) that you wish to answer from a statistical perspective. There are two types of questions to pick from, the first where you want to compare a *feature* of a specific group to that of the general population, the second where you wish to compare two populations that possess the same feature. The first type of question is called a one-sample problem, while the latter is called a two-sample problem.

An example of a one-sample problem might be, "Do Olin students sleep less than other college students in general?" Here, *amount of sleep per night* is the feature, which is a continuous random variable that has some underlying probability distribution. A typical two-sample problem might be "Do male and female Olin students have the same average height?" In this problem, the feature is *height*, which is also a continuous random variable.

Pick a question that you find interesting, as well as one for which you can collect data. Two-sample problems will tend to be easier to answer (especially for this particular project) because one does not need to obtain any information regarding the population statistical parameters of the populations being studied. In the one-sample case, you would need to know the values for μ and σ, as well as the probability distribution, for the larger population being compared. If you decide to do a one-sample type of problem, use the web to find repositories of large-scale statistical information. Good places to look at would be the US Census databases `http://www.census.gov/` or the Survey Documentation and Analysis site at UC Berkeley `http://sda.berkeley.edu/cgi-bin/hsda?harcsda+gss06`.

2.2 The Data

After formulating your question, you will need to collect data to test your hypothesis. This will be in the form of a single random sample of size n (for a one-sample problem), or two random samples of sizes n_1 and n_2 in the two-sample case. Be sure you can collect such data before deciding on your question, as you will only be able to test your hypothesis if you are able to collect this sample data.

After collecting the data, you will need to summarize it using numerical and graphical methods. Examples of numerical methods would be to report the sample statistical parameters such as the mean $\bar{x}$ and standard deviation s. The standard graphical way to represent your data would be a relative frequency histogram (see Sec. 3.1 of your book). Microsoft Excel has a nice and easy

to use Data Analysis Pack that provides tools for creating such plots. One simply has to define appropriate bins for their random variable x and Excel does the rest for you!

Some questions that you need to answer would include:

- Is my random variable discrete or continuous? What are its range of values?
- Looking at the shape of the relative frequency histogram for x, what probability distribution does my random variable follow (normal, binomial, etc.)?
- What is the shape of the histogram like? Is it unimodal, bimodal, or multimodal? Is it symmetric or skewed to the left or right?
- What are my computed measures of central tendency (sample mean and median) and measures of dispersion (sample standard deviation, range = max - min)?

2.3 Confidence Intervals

Next, compute the appropriate confidence interval(s) (CI) based on the question you first posed. In the one-sample case, you could be testing to see if the mean amount of sleep Olin students get (μ) is equal to that of the population of students in general (μ_0). So determine a CI that contains μ with some appropriate level of significance. On the other hand, in a two-sample problem, you might be testing to see if the mean height of Olin men (μ_1) is the same as Olin women (μ_2), so find a CI that contains the difference of means $\mu_1 - \mu_2$. You might instead be testing variation instead of means, and so you would use the appropriate CI based on your statistic for either σ^2 in the one-sample case or σ_1^2/σ_2^2 in the two-sample case.

2.4 Hypothesis Tests

You also clearly have to state your statistical hypothesis and which test(s) you intend to use in order to answer your question.

Some questions that you need to answer would include:

- What is my null hypothesis H_0?
- What is my alternate hypothesis H_1? Is it one- or two-sided?
- What statistic(s) will I use to test my hypothesis (z, t, χ^2, F, ...)
- What is my rejection region for my test? Does the computed value of my test statistic lie in the rejection region?
- Based on the evidence, am I able to reject the null hypothesis, or do I fail to reject it?

3 Analysis of Results

The most important part of your project comes when you analyze the results that you have tabulated. We do not want you to simply compile a bunch of graphs and measures and put them into a document. Examine all the evidence that you have found and comment on how this information allows you to answer your question from a statistical standpoint.

For example, if I were conducting an analysis on the mean heights of female and male Olin students, I would first comment on the shapes of the relative frequency histograms for each group

and note any important features and explain why the shape allows me to assume that the underlying populations follow a particular probability distribution. I would also comment on the numerical values of the sample statistical parameters (e.g., $\bar{x_1}$, $\bar{x_2}$, s_1, and s_2).

Next, I would explain why I chose to use a particular equation for the CI for the difference of means $\mu_1 - \mu_2$, as well as explain how the values of the endpoints of the interval allow me to conclude something about the question I originally posed. Finally, I would look at the results of my hypothesis, particularly the value of the test statistic in relation to the rejection region, and using this information, explain my answer (from a statistical perspective) to the original question I posed.