IE 300 Project 2b – Central Limit Theorem

For lab/discussion sections, week of Nov. 6-10 and Nov. 13-17

Preliminaries

Project 2b is to be completed by the sam student teams as Project 2a, with each team containing two students (excepting one or two teams that may contain 3 students). Teams are assigned by the instructor. Teams are identified by student last names.

Real Data Analysis

In this part of the project we will consider the Bluegill (a fish species) catch data from Project 2a, from McDermott Lake in Wisconsin over 15 different days in summer of 2017, and now we will add a second set of Bluegill catch data from Sandy Beach Lake in Wisconsin over 19 days in summer of 2017 (note one date has two entries).

Specifically, you will analyze the data in the files

Isermann_Bluegill_Sampling_TotalLengths-1.csv,

Isermann_Bluegill_sampling_catch_1.csv

Isermann_Bluegill_Sampling_TotalLengths-2.csv, and

Isermann_Bluegill_sampling_catch_2.csv

using appropriate tools and techniques. The *catch* data represents the number of Bluegill caught on the given date, and the *TotalLengths* represents the lengths of each Bluegill caught in units of mm. **Note:** as before this is all real data so won't fit models as clearly as simulated data. For this part of Project 2, we will work only with the TotalLengths data sets.

1. For each of the individual TotalLengths data sets, analyze the data for Normality using (1) visual analyses based on histograms and (2) normality plots. Discuss your findings.

2. Compute the following sample statistics for each of the two TotalLengths data sets:
   Sample mean
   Sample variance
   Sample standard deviation
   Sample median

3. For each of the individual TotalLengths data sets, partition the data into subsets with number of entries around 25. You should choose plus or minus one or two on the data subset sizes here, so that the partitions are even and have the smallest number of "leftovers" as possible. The size of the data subsets should be different for TotalLengths-1 (n_1), and TotalLengths-2 (n_2). We will ignore the "leftovers" going forward and work just with the evenly sized data subsets. Clearly state the sizes of the data subsets for TotalLengths-1 and TotalLengths-2, i.e., what is n_1 and n_2, and how many "leftovers" were there for each?

4. We will now create two new data sets, where the entries are the **subset means**. Compute the means of all data subsets (where these data subsets are sizes n_1 and n_2, respectively) and store as new files titled MeanLengths-1 and MeanLengths-2.

5. Assess both MeanLengths data sets for Normality, again using (1) visual analyses based on histograms and (2) normality plots. Discuss your findings and compare to your results in 1. (E.g., does the means data appear more or less Normal than the original data?)

6. Consider the following known result for differences in sample means:
   If we have two independent populations with means μ_1 and μ_2, and finite variances σ_1^2 and σ_2^2, and if $\bar{X}_1$ and $\bar{X}_2$ are the sample means of two independent random samples of sizes n_1 and n_2 from these populations, then the distribution of the random variable Z defined as

$$Z := \frac{\bar{X}_1 - \bar{X}_2 - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

   is approximately Standard Normal (assuming the usual conditions of the CLT apply). If the two underlying populations are Normally distributed, then Z is exactly Standard Normal.

   Based on your preceding analyses, discuss the Normality of the standardized difference in sample means for the Bluegill data sets. What do you expect and why? Can you verify using the data you have?

Turn in the following:

1. A concise but complete report including your analysis and support for your conclusions.
2. Relevant histograms, graphs and/or plots.
3. Include references to routines used or code for any routines you wrote yourselves in an appendix.