Stat 359 / Bio 550B Assignment 2

1. (14 marks) Use R to take 10, 100, 1000 samples of size 10, 30, 100 from 2 distributions.

   1) t distribution with df=3

   3) Binomial(n=1, p=0.85)

   (a) (8 marks) Investigate how the CLT works when sampling from the above distributions and for each sample size.

   • Write an R function to simulate the distribution fo the sample mean. The R function has three arguments, the number of samples (N), the sample size (n), and the distribution. In the function, you generate N samples from the given distribution and each sample has size n. Then you compute N sample means. The output of the function includes a histogram of the sample means and a normality test for the sample means. For the normality test, use the function shapiro.test() and read up on the Shapiro-Wilk test for normality in the help files and in your text.

   • Use histograms to illustrate the distributions of the sample mean for each sample size, number of samples, and distribution.

   • Organize the 18 (3x3x2) histograms into 2 pages for your solution.

   (b) (2 marks) What do you notice as sample sizes increase?

   (c) (2 marks) What do you notice as the number of samples increases?

   (d) (2 marks) How does the distribution affect the outcome?

2. (10 marks) A mixture of salt and sucrose was tasted to investigate how saltiness was judged depending on sucrose concentration and the data are contained in the file

   `salt.txt`

   (a) (2 marks) Determine whether the data come from a symmetric distribution. Comment.

   (b) (4 marks) Develop a bootstrap test to determine if the mean and median are equal. What do you conclude?

   (c) (2 marks) Estimate the skew (γ_1) and kurtosis (γ_2) of this distribution using the data.

   (d) (2 marks) Based on your analysis above, what do you conclude about the distribution?

3. (12 marks) Data on the per diem fecundity (fecundity.txt) (number of eggs laid per female per day for the first 14 days of life) for 25 females on 2 genetic lines of the fruit fly *Drosophila melanogaster* are provided. Resistant (RS) to DDT were selectively bred and non-selected (NS) was the control.

   (a) (2 marks) Construct a normal QQplot for the RS line and comment on the distribution of RS.

(b) (4 marks) Compute the variances of the RS and NS lines. Construct side-by-side boxplots for RS and NS lines. Perform a hypothesis test to determine if RS and DS lines differ in population variance and provide the output from this test. Comment on the equal variance assumption for RS and NS lines.

(c) (3 marks) Do RS and NS lines differ in population mean fecundity? Comment on your results.

(d) (3 marks) Perform a nonparametric Wilcoxon test to see whether RS and NS lines differ in population mean fecundity.

4. (10 marks) Fusible interlinings are being used with increasing frequency to support outer fabrics and improve the shape and drape of various pieces of clothing. The data on extensibility (100%) at 100 gm/cm for both high quality fabric (H) and poor -quality fabric (P) specimens is given in *fabric.txt*.

(a) (4 marks) Construct normal qq plots to verify the plausibility of both samples having been selected from normal population distributions. Comment.

(b) (4 marks) Construct a comparative boxplot. Does it suggest that there is a difference between true average extensibility for high-quality fabric specimens and that for poor-quality specimens?

(c) (2 marks) Determine whether true average extensibility differs for the two types of fabric. Comment.

Graduate students only:

5. (5 marks) Create a bootstrapping function to complete tests in question 3 and 4. How do these results compare with your results from above?