**Applied Statistics and Design of Experiments**

**Exercise 1**

**January 9, 2019**

Questions 1-4 will explore how well we can assess whether a sample of data comes from a normal distribution.

1. Take a random sample of size 10 from the standard normal distribution (mean 0, sd 1). Make a histogram and q-q plot of your sample. Does the distribution clearly look like a normal distribution? Does it clearly look non-normal? Repeat the sampling procedure at least 20 times. Do the samples consistently appear to be normal or non-normal?

2. One way to model data from a heavy-tailed distribution is to use a t distribution. The family of t distributions is indexed by a parameter called the degrees of freedom. For this question we will use the t distribution with 3 degrees of freedom. We will denote this distribution as t(3). Make a plot of the pdf of the t distribution with 3 degrees of freedom. Add the standard normal distribution to the plot. Describe the differences between the two distributions.

3. Repeat question 1 using samples from the t(3) distribution. Compare results with the results of question 1.

4. Repeat questions 1 and 3 with a sample size of 100. How do the results differ? How does the ability to detect non-normality depend on sample size?

Questions 5-8 will use simulation studies to explore the validity of the standard error formula for describing the variability of a sample mean.

5. Perform a simulation study to assess the sampling distribution of the sample mean for sampling from the standard normal distribution with sample size 10. To do this follow the steps below:

(i) generate 200 independent random samples from the standard normal distribution;

(ii) for each of the 200 samples calculate the sample mean;

(iii) make a histogram of the 200 sample means;

(iv) calculate the mean, variance and SD of the sample means.

How well do the mean, variance and SD of the sample means compare to the values you would expect from the theoretical SE formula?

6. Repeat the simulation study using a sample of size 40.

7. Using the SE formula, what is the minimum sample size required in order for the sample mean to have SD of 0.1 or less when sampling from the standard normal distribution? Perform a simulation study to verify your answer.

8. Consider sampling from a Bernoulli distribution with success probability p=0.1. The sample means are now sample proportions. What is the minimum sample size required to have a sample proportion with SD of 0.01 or less? Verify using a simulation study.

Questions 9-13 will use simulation studies to explore the Central Limit Theorem.

9. Perform a simulation study to assess the sampling distribution of the sample mean for sampling from the exponential distribution with mean equal to 3, with sample size 10. Does the distribution of the sample mean appear to be closely approximated by a normal distribution? Use a histogram and q-q plot.

10. How well do the mean, variance and SD of the sample means compare to the values you would expect from theory?

11. Repeat questions 5 and 6 using a sample of size 40.

12. Now consider sampling from the Bernoulli distribution with success probability p=0.1. What is the minimum sample size required in order that the sample proportion has an approximately normal distribution? (Note: this question is somewhat subjective; there is no precise answer.)

13. How would the answer to question 12 change if the success probability was p=0.5 instead of 0.1?

Questions 14-16 will explore the calculation and properties of confidence intervals. We will use the built-in R data set ‘morley’, which contains measurements of the speed of light (in km/sec, with 299000 subtracted) performed in 5 experiments with 20 runs each. We will only use the 20 measurements from Experiment 1 for this assignment. We will consider these 20 measurements to be a random sample from a population with population mean equal to the true value of 792 (the true speed of light is 299,792 km/sec).

14. Calculate a 95% confidence interval for the speed of light using the 20 measurements from Experiment 1.

15. Is the true value for the speed of light in the 95% confidence interval? What does this say about the bias or lack of bias of the experimental measurement technique.

16. Perform a simulation study to assess the performance of the confidence interval for the speed of light based on 20 measurements. For the simulations, treat the Experiment 1 sample as if it were the population and sample with replacement from this “population”. Use a sample size of 20, i.e., the same as the actual sample size. Generate 1000 samples of size 20 from the “population” and calculate the confidence interval for each sample. What percentage of the confidence intervals contain the true speed of light? What percentage contains the “population” mean? Explain.