COSC 6323 – Statistical Methods in Research

Homework 1

January 24, 2020

Problems

1. **Q1.txt** shows the times in days from remission induction to relapse for 51 patients with acute nonlymphoblastic leukemia who were treated on a common protocol at university and private institutions in the Pacific Northwest. This is a portion of a larger study reported by Glucksberg *et al.* (1981).

   (a) What is the relation between the mean and the median? What does this mean about the shape of the distribution? Do the box plot support this?

   (b) Identify any outliers in this data set. Can you think of any reasons for these outliers? Can we just "throw them away"? Note that the mean time of remission is 292.39 days and the median time is 249.

   (c) Approximately what percent of these patients were in remission for less than one year?

2. A lottery that sells 150,000 tickets has the following prize structure:
   (a) First prize of $50,000
   (b) 5 second prizes of $10,000
   (c) 25 third prizes of $1000
   (d) 1000 fourth prizes of $10

   i. Let Y be the winning amount of a randomly drawn lottery ticket. Describe the probability distribution of Y.

   ii. Compute the mean expected value of the ticket.

   iii. If the ticket costs $1.00, is the purchase of the ticket worthwhile? Explain your answer.

   iv. Compute the standard deviation of this distribution. Comment on the usefulness of the standard deviation as a measure of dispersion for this distribution.

3. Based on data from the 2007 National health Interview Survey, it is estimated that "10% of adults experienced feelings of sadness for all, most, or some of the time" during the 30 days prior to the interview. You interview a random sample of 68 people who have recently filed for unemployment benefits in your county, and ask this same question in your survey.

(a) Identify the implied target population for your study.
(b) If the proportion of your population with these feelings is the same as the 10% nationally, what is the probability that your sample will have 12 or more people with these feelings?
(c) What would you conclude if your sample did have 12 or more people with these feelings?
(d) If the true percentage of your population with these feelings is 10%, what is the probability that the sample percentage will differ from this by more than 5%?

4. The family incomes in a certain city in 1970 had a mean of $14,200 with a standard deviation of $2600. A random sample of 75 families taken in 1975 produced $\bar{y}$= $15, 300 (adjusted for inflation).

   (a) Assume σ has remained unchanged and test to see whether mean income has changed using a 0.05 level of significance.
   (b) Construct a 99% confidence interval on mean family income in 1975.
   (c) Construct the power curve for the test in part (a).

5. A large national survey of American dietary habits showed a mean calorie consumption of 2700 kcal and a standard deviation of 450 kcal among teenage boys. You are studying dietary habits of children in your county to see if they differ from the national norm.
   (a) In a sample of 36 teenage boys, you find a mean consumption of 2620 kcal. At α = 0.05, is this significant evidence that the mean in your county differs from the national mean? Assume that the standard deviation observed nationally can be used for σ .
   (b) Using α=0.05 and sample of size 36, what is the probability that you will actually be able to detect a situation where your county has a mean of only 2600 kcal? (That is, what is the power if μ=2600?)

Submission

- Zip your submission folder (contains code, documents, and plots) and submit to blackboard. Name your zip file "**`FirstName_LastName_HW1.zip`**". One submission per individual.
- The homework should be submitted on BlackBoard by **4pm on February 7**, **2019**.
- The homework report should be a pdf generated in Latex with figures from R only. All the figures and plots should have appropriate labels, titles and annotations.
- The code should also be provided.

Grading

Total points **25 pts** (5 pts each)