OPIM 5603-B14 — Statistics in Business Analytics
Fall 2019, University of Connecticut

Homework 7 - v1

Instructions: Please complete the following questions and submit them as an RNotebook (as an `Rmd` file) via the submission link on HuskyCT. You must submit the assignment by the time and due date listed on the course syllabus. Failure to submit a file by the deadline will result in a score of 0 on the assignment.

Set the heading of the RNotebook as an `html_document`, with a table of contents and without numbered sections. Add your name and a date to the header as well. The solution to each problem should be a separate section (specified by #), and each subproblem should be set as a subsection (specified as ##). For example, for Problem 2, you should have a section titled Problem 2, specified by:

```
# Problem 2
```

in your RNotebook. Also, for subproblem b in Problem 2, you should have a subsection, specified by:

```
## Problem 2b
```

As with all course material, the problems appearing in this homework assignment are taken from the instructor's real-world experiences, from other courses taught at the University of Connecticut, and from the sources listed in the course syllabus.

Note that R code submitted should work independent of the data that sits in the data structure. For example, suppose there was a vector `r_vec` with the values $(1, 2, 6)$ and the problem asks for you to create R code to create a vector `answer` which doubles each element of `r_vec`. The answer

$$\texttt{answer} \leftarrow \text{c}(2, 4, 12)$$

would be given no credit. The answer

$$\texttt{answer} \leftarrow 2^{*}\texttt{r\_vec}$$

would be an appropriate answer.

You must show all steps in your solution. For example, if a problem asks for the expected value of a random variable that is binomially distributed with $n = 10$ and $\pi = 0.3$, and you simply write

$$3,$$

this will be given no credit. However,

$$10 * 0.3$$

would be given credit.

If you have any questions, please submit them via email to the instructor and/or the teaching assistant prior to submitting your solution.

Problem 1 (25 points)

Let X be a random variable that is normally distributed with mean 10 and standard deviation 2. Let Y be a random variable that is normally distributed with mean 15 and standard deviation 1. Let Z be a random variable that is $2 \cdot X + 3 \cdot Y$.

a. What is the expected value of X? Answer this question without simulation.

b. What is the expected value of Y? Answer this question without simulation.

c. What is the expected value of Z? Answer this question without simulation.

d. Use simulation to estimate the correlation of Y and Z. What is your estimate of the correlation?

e. Using your estimate from the previous part, and without additional simulation, what is the variance of $Y + Z$?

f. Simulate $Y + Z$. What is the estimated variance of $Y + Z$ based on the simulation?

Problem 2 (25 points)

In class we discussed Chebyshev's rule, which states that regardless of how the data are distributed, at least $(1 - 1/k^2) \cdot 100\%$ of the distribution of a variable will fall within k standard deviations of the mean (for $k > 1$).

a. For $k = 2$ and for $k = 3$, what is the maximum amount of the distribution that will fall within k standard deviations of the mean?

b. Find a distribution that minimizes the portion of the distribution (mass for discrete, density for continuous) that falls within 2 standard deviations of the mean. How does this compare with Chebyshev's rule? Explain.

c. Find a distribution that minimizes the portion of the distribution (mass for discrete, density for continuous) that falls within 3 standard deviations of the mean. How does this compare with Chebyshev's rule? Explain.

Problem 3 (25 points)

Suppose that the true distribution of the median salary in an affluent suburb of New York is normally distributed with mean $120,000 and standard deviation $10,000.

a. Suppose I take a random sample of 10 individuals from the area. What is the distribution of the mean salary in the sample? Please indicate the parameters of the distribution as well as the distribution type.

b. Simulate 1000 samples of size 10. What is the average value of the sample mean? What is the standard deviation of the sample mean? How does that compare with your answer to the previous part.

c. Simulate 10000 samples of size 10. What is the average value of the sample mean? What is the standard deviation of the sample mean? How does that compare with your answer to the previous parts.

d. Simulate 100000 samples of size 10. What is the average value of the sample mean? What is the standard deviation of the sample mean? How does that compare with your answer to the previous parts.

e. Now suppose you take a random sample of size 30. What is the distribution of the mean salary in the sample? Please indicate the parameters of the distribution as well as the distribution type.

f. Simulate 100000 samples of size 30. What is the average value of the sample mean? What is the standard deviation of the sample mean? How does that compare with your answer to the previous parts.

Problem 4 (25 points)

You are looking to estimate the proportion of pet owners that have pet insurance. You know from experience that at least 10% have pet insurance. A statistician takes a random sample of 1000 pet owners and creates a 95% confidence interval for the proportion of pet owners that have pet insurance. The confidence interval is $[0.23, 0.31]$. For each part below, if you can answer the question, provide the answer and an explanation. If you cannot answer the question, provide details on why you cannot.

a. What is the probability that the true population proportion of pet owners that have pet insurance is above 0.1?

b. What is the probability that the true population proportion of pet owners that have pet insurance is between $[0.23, 0.31]$?

c. David Bergman has decided to create a procedure for building interval estimates. He takes a random sample of 100 pet owners. If the proportion of pet owners in the sample that have pet insurance is less than 0.1, he reports an interval estimate of $[0.0, 0.05]$. If the proportion of pet owners in the sample that have pet insurance is greater than or equal to 0.1, he reports an interval estimate of $[0.1, 1.0]$. Ignoring the statistician's test, but taking the rest of the problem statement into account, what is the maximum probability that David Bergman's test reports an interval of $[0.0, 0.05]$?

d. Suppose the true proportion of per owners that have pet insurance is 0.3. What is the confidence level of David Bergman's test?