Homework 2

P6104 Fall 2023

Due: Sunday, October 8 (11:59 p.m.)

Please do all problems below showing all work unless you are explicitly told to use R. You can use R to check your solutions for those problems that you do by hand if you’d like but we would like to see the steps of each solution clearly explained.

**Module 04 Questions**

1. Assume that the population of human body temperatures is normal and has a mean of = 98ºF, as is commonly believed. Also assume that the population standard deviation is 0.62ºF (based on the data from a researcher in the department of Biology, BSU). A sample of size *n* = 35 is randomly selected.

1. Find the probability of getting a sample mean of 98.2ºF or lower.
2. Find the probability that a randomly selected individual has a body temperature of 98.2ºF or lower.
3. Find the probability of getting a sample mean within 0.15 ºF of the population mean.
4. Find the probability of getting a sample mean that will differ from the population mean  by more than 0.2ºF.
5. Use R to compute the probability of observing a sample mean that is greater than 98.7ºF. Include the one line of R code that you used to answer this question.

2. We want to estimate the mean of a population. A random sample of subjects is selected and the sample mean is computed. What is the probability that the sample mean is within 3 units of the true mean if the standard error is 1.8?

3. It is believed that 7.8% of children are diabetic. Lack of physical activity in children is considered to be a risk factor for diabetes. In particular, children who engage in less than 6 hours per week are considered “at risk.” Among diabetic children, the amount of weekly activity is approximately normally distributed with a mean of 4.5 hours per week and a standard deviation of 0.75 hours. Among children who are not diabetic, the amount of weekly activity is approximately normally distributed with a mean of 8.7 hours per week and a standard deviation of 2.2 hours.

1. For children who are not diabetic, what are the values that cut off the middle 80% of physical activity? Provide a clear interpretation of these values.
2. What is the probability that a child develops diabetes given that he is “at risk”?

4. The distribution of survival times of cancer patients after treatment is typically strongly right skewed. A medical study on guinea pigs was conducted to investigate the survival time after treatment for tumors. The data for this study can be found in the file “gpigs.txt” which gives the survival time in days for the sample of 72 guinea pigs.

1. Make a histogram of the survival times. Describe the shape of the distribution of the individual values.
2. The CLT suggests that the sampling distribution of the sample mean based on 72 survival times is approximately normal even though the distribution of an individual’s survival time is not. Use bootstrapping with B = 1000 bootstrap samples to either confirm or deny that the sampling distribution of the sample mean based on 72 survival times is approximately normal. Briefly explain your conclusion and also compute and report the bootstrap standard error of the sample mean. Use a seed value of 54321.
3. The CLT says that the standard error of the sample mean is given by . Since we do not know the true value of , we can estimate it with the sample standard deviation. Is the bootstrap standard error of the mean computed in part (b) close to the standard error based on the CLT?
4. The file “gpigs20.txt” contains a simple random sample of 20 of the 72 survival times in the “gpigs.txt” file. Create and inspect a bootstrap distribution for the sample mean based on 20 survival times using the data in the “gpigs20.txt” file. Use bootstrapping with B = 1000 bootstrap samples to construct a bootstrap distribution for the sample mean based on 20 survival times. Use a seed value of 12345. Does the bootstrap distribution suggest that the sampling distribution for the sample mean based on 20 survival times is closer to normal or farther from normal than the sampling distribution of the sample mean based on 72 survival times? Also compute and report the bootstrap standard error of the sample mean.
5. In a few sentences, explain why it makes sense that the bootstrap standard error that you computed in (d) is larger than the bootstrap standard error you computed in part (b).

**Module 05 Questions**

5. Increases in worker injuries and disability claims have prompted renewed interest in workplace design and regulation. As one particular aspect of this, employees required to do regular lifting should not have to handle unsafe loads. One study reported information regarding a random sample of 18 male postal workers. The sample mean rating of acceptable load attained with a work-simulating test was found to be kg. Suppose we know that the standard deviation was = 4.3 kg. Suppose that in a population of all male postal workers, the distribution of rating of acceptable load can be modeled approximately using a normal distribution with mean .

1. Construct a 95% confidence interval for .
2. Give a practical interpretation and a probability interpretation of the 95% confidence interval in the context of the problem.
3. Holding all else constant, how would the width of the confidence interval change if the sample size increased to 49 postal workers? Explain.
4. Holding all else constant, how would the width of the confidence interval change if the confidence level were to increase to 99%? Explain.
5. Holding all else constant, how would the width of the confidence interval change if the standard deviation were 2.5 kg rather than 4.3 kg? Explain.
6. Is the population mean acceptable load for all male postal workers different from 12kg? Conduct a hypothesis test using a 5% significance level. Be sure to include the following:
   1. Define the parameter of interest.
   2. State the null and alternative hypotheses.
   3. State the significance level.
   4. Provide a check for any necessary assumptions.
   5. State your test statistic and its null distribution.
   6. Compute the test statistic BY HAND. Make a decision to reject or fail to reject based on either a p-value or by comparing the test statistic to a critical value (you can choose whichever you prefer).
   7. State your conclusion in the context of the problem.
7. In your own words, explain how you could have used your confidence interval from part (a) to answer the question in part (f).

6. Researchers want to estimate the mean diastolic blood pressure level of adult females in a small village. From large population studies, they know that the variance of the diastolic blood pressure level of adult females is 126 (mmHG)2. How many adult female villagers must they randomly select and test if they want to be 90% confident that the true mean diastolic blood pressure level of adult females in the village is within 4 mmHG of the sample mean?

7. Medical personnel are required to report suspected cases of child abuse. Because some diseases have symptoms that mimic those of child abuse, doctors who see a child with these symptoms must decide between two competing hypotheses:

H0: symptoms are due to child abuse

H1: symptoms are due to disease

(Although these are not hypotheses about a population characteristic, this problem illustrates the definitions of Type I and Type II errors.) The article “Blurred Line Between Illness, Abuse Creates Problem for Authorities” (*Macon Telegraph,* February 28, 2000) included the following quote from a doctor in Atlanta regarding the consequences of making an incorrect decision: “If it’s disease, the worst you have is an angry family. If it is abuse, the other kids (in the family) are in deadly danger.”

1. For the given hypotheses, describe Type I and Type II errors.
2. Based on the quote regarding consequences of the two kinds of error, which type of error does the doctor quoted consider more serious? Explain.