SOC6302 Winter 2023

Assignment 2

Due date: Friday 10 March, 11:59pm

Details

There are **100 points** in total.

You will need to submit both your answers to the questions and accompanying R code. You should submit:

- your Quarto file; and
- the knitted PDF resulting from your Quarto file.

Please submit both files via Quercus.

Remember to:

- Label the answers to each question
- Label any graphs clearly with suitable axis labels and titles
- Comment your code so that it is easy to understand

Question 1 (40 points)

This question relates to the birthweight dataset, which lists the birth weight (in grams) of sample of 500 babies.

a) Plot a histogram of the birthweights and comment briefly.

b) Calculate the mean, median, standard deviation, and range of the birthweights.

c) Assume that the sample of 500 babies are drawn from a Normal distribution with mean and standard deviation of what you calculated in part b).

   i) What is the probability of observing a baby that has low birthweight (formally defined as less than 2500g)?
   ii) What is the probability of observing a baby that has a birthweight of more than 3700g?
   iii) What is the probability of observing a birthweight between 3400g and 3500g?

d) i) How many standard deviations above the mean is a baby that weighs 4000g at birth?
   ii) We are told a baby has a Z score of 0.13. What is the baby's weight?

e) Suppose we observe 15 more birthweights (listed below). Discuss briefly with the aid of calculations and/or graphs whether you think these 15 babies are likely to be sampled from the same population as above.

birthweight
3984
3913
3864
3620
4117
4215
3857
3935
3886
4116
4110
4080
3897
4021
4243

Question 2 (20 points)

The following picture is a sheepdog that looks like a mop.

You are interested in seeing what Toronto residents think this is a picture of. You randomly sample 100 residents, show them the picture for 2 seconds, and then ask them whether they think it is a picture of a sheepdog or a mop.

a) Assume the true proportion of people believing the picture is a sheepdog is 0.8. Simulate your data as specified above. Set the seed to be 675 (i.e. before simulating data, please type `set.seed(675)`). What is the observed proportion of people who think the picture is a sheep dog?
b) Calculate and interpret a 95% confidence interval for the observed proportion.
c) Calculate and interpret a 90% confidence interval for the observed proportion.

Question 3 (40 points)

My 18-month old seems quite tall for his age – he is about 87cm. I decide to investigate heights of other toddlers the same age to see if he is indeed taller than average.

Suppose the true mean height of 18-month olds μ is 80cm and the true standard deviation σ is 3. I do an experiment take a random sample of 20 toddlers and record their height. I repeat this experiment 10 times, so I have 10 samples of 20 heights.

a) Simulate these data in R, assuming the sample is drawn from a normal distribution with the true values of μ and σ. Before simulating data, please type `set.seed(853)`.
b) Calculate the mean height recorded in each of the 10 experiments, and plot these mean heights as a histogram.
c) Repeat b) and c) but now instead of 10 experiments, assume I have 1000 experiments (you will need to set the seed again). Interpret the difference in the histograms.

Now suppose I sample 10 toddlers and observe a mean height of 80.5cm and a standard deviation of 2.4cm.

d) Construct a two-tailed hypothesis test to test whether my toddler's height is different from the general population of toddlers. Would the null hypothesis be rejected at the 5% level? For this question you need to define your hypotheses, do the appropriate calculations, and interpret your results.

Review Questions (not to hand in)

- True or False: the range is always larger than the interquartile range.
- Is the mean always equal to the median? If not, when are they equal?
- I have a dataset of marathon times of people who completed the New York marathon. In the dataset I also have information on: age, hours trained, and whether or not the runner had completed the New York marathon previously. Describe three graphs and/or summary measures to explore patterns in this dataset.
- In a standard set of playing cards, what's the probability of drawing a red card? What's the probability of drawing a Queen? Given I know the card is a black card, what's the probability that it's a club?
- I flip a coin twice. What's the probability of seeing two heads? What's the probability of seeing at least one head?
- Which is larger, a 95% confidence interval or a 99% confidence interval?
- True or False: In a standard normal distribution, the probability of being above 3.5 standard deviations above the mean is equal to the probability of being 2.5 standard deviations below the mean.
- I like knitting and I am 36 years old. Pretend I know the distribution of ages of people who knit in Australia versus people who knit in Canada. The standard deviation of both distributions is the same. In Australia my Z score is -2.5. In Canada my Z score is -1. Is the mean age of knitters older in Australia or Canada?
- I like knitting and I am 36 years old. Pretend I know the distribution of ages of people who knit in Australia versus people who knit in Canada. The mean of both distributions is the same. In Australia my Z score is -2.5. In Canada my Z score is -1. Is the standard deviation age of knitters larger in Australia or Canada?