Assignment 5 - Inference for a Population Mean. Due October 14, 11:59pm 2018

EPIB607 - Inferential Statistics[a]

[a]Fall 2018, McGill University

This version was compiled on October 7, 2018

In this assignment you will practice conducting inference for a one sample mean using either the z procedure, t procedure, or the bootstrap. Answers should be given in full sentences (DO NOT just provide the number). All figures should have appropriately labeled axes, titles and captions (if necessary). All graphs and calculations are to be completed in an R Markdown document using the provided template. You are free to choose any function from any package to complete the assignment. Concise answers will be rewarded. Be brief and to the point. Please submit both the compiled HTML report and the source file (.Rmd) to myCourses by October 14, 2018, 11:59pm. Both HTML and .Rmd files should be saved as 'IDnumber_LastName_FirstName_EPIB607_A5'.

t-test | One sample mean | Bootstrap

Template

The `.Rmd` template for Assignment 5 is available here

1. Food intake and weight gain

If we increase our food intake, we generally gain weight. Nutrition scientists can calculate the amount of weight gain that would be associated with a given increase in calories. In one study, 16 nonobese adults, aged 25 to 36 years, were fed 1000 calories per day in excess of the calories needed to maintain a stable body weight. The subjects maintained this diet for 8 weeks, so they consumed a total of 56,000 extra calories. According to theory, 3500 extra calories will translate into a weight gain of 1 pound. Therfore we expect each of these subjects to gain 56,000/3500=16 pounds (lb). Here are the weights (given in the `weightgain.csv` file) before and after the 8-week period expressed in kilograms (kg):

```
weight <- read.csv("weightgain.csv")
```

```
#     subject before after
#  1        1   55.7  61.7
#  2        2   54.9  58.8
#  3        3   59.6  66.0
#  4        4   62.3  66.2
#  5        5   74.2  79.0
#  6        6   75.6  82.3
#  7        7   70.7  74.3
#  8        8   53.3  59.3
#  9        9   73.3  79.1
#  10      10   63.4  66.0
#  11      11   68.1  73.4
#  12      12   73.7  76.9
#  13      13   91.7  93.1
#  14      14   55.9  63.0
#  15      15   61.7  68.2
#  16      16   57.8  60.3
```

a. Calculate a 95% confidence interval for the mean weight change and give a sentence explaining the meaning of the 95%. State your assumptions.
b. Calculate a 95% bootstrap confidence interval for the mean weight change and compare it to the one obtained in part (a). Comment on the bootstrap sampling distribution and compare it to the assumptions you made in part (a).
c. Convert the units of the mean weight gain and 95% confidence interval to pounds. Note that 1 kilogram is equal to 2.2 pounds.
d. Test the null hypothesis that the mean weight gain is 16 lbs. State your assumptions and justify your choice of test. Be sure to specify the null and alternative hypotheses. What do you conclude?

2. Attitudes toward school

The Survey of Study Habits and Attitudes (SSHA) is a psychological test that measures the motivation, attitude toward school, and study habits of students. Scores range from 0 to 200. The mean score for U.S. college students is about 115, and the standard deviation is about 30. A teacher who suspects that older students have better attitudes toward school gives the SSHA to 25 students who are at least 30 years of age. Their mean score is $\bar{y} = 132.2$ with a sample standard deviation $s = 28$.

a. The teacher asks you to carry out a formal statistical test for her hypothesis. Perform a test, provide a 95% confidence interval and state your conclusion clearly.
b. What assumptions did you use in part (a). Which of these assumptions is most important to the validity of your conclusion in part (a).

3. Does a full moon affect behavior?

Many people believe that the moon influences the actions of some individuals. A study of dementia patients in nursing homes recorded various types of disruptive behaviors every day for 12 weeks. Days were classified as moon days if they were in a 3-day period centered at the day of the full moon. For each patient, the average number of disruptive behaviors was computed for moon days and for all other days. The hypothesis is that moon days will lead to more disruptive behavior. We look at a data set consisting of observations on 15 dementia patients in nursing homes (available in the `fullmoon.csv` file):

```
fullmoon <- read.csv("fullmoon.csv")
```

```
#     patient moon_days other_days
#  1        1      3.33       0.27
#  2        2      3.67       0.59
#  3        3      2.67       0.32
#  4        4      3.33       0.19
#  5        5      3.33       1.26
#  6        6      3.67       0.11
#  7        7      4.67       0.30
#  8        8      2.67       0.40
#  9        9      6.00       1.59
#  10      10      4.33       0.60
#  11      11      3.33       0.65
#  12      12      0.67       0.69
#  13      13      1.33       1.26
#  14      14      0.33       0.23
#  15      15      2.00       0.38
```

a. Calculate a 95% confidence interval for the mean difference in disruptive behaviors. State the assumptions you used to calculate this interval.
b. Calculate a 95% bootstrap confidence interval for the mean difference in disruptive behaviors and compare to the one obtained in part (a). Comment on the bootstrap sampling distribution and compare it to the assumptions you made in part (a).
c. Test the hypothesis that moon days will lead to more disruptive behavior. State your assumptions and provide a brief conclusion based on your analysis.
d. Find the minimum value of the mean difference in disruptive behaviors ($\bar{y}$) needed to reject the null hypothesis.
e. What is the probability of detecting an increase of 1.0 aggressive behavior per day during moon days? *Hint: calculate the probability of the event calculated in part (d) using a normal distribution with $\mu = 1$ and $\sigma =$ the standard error of the mean*

4. How deep is the ocean?

This question is based on the in-class Exercise on sampling distributions and builds on Question 4 from Assignment 4. For your sample of $n = 20$ of depths of the ocean

a. Calculate a 95% Confidence interval using the t procedure
b. Plot the qnorm, bootstrap, and t procedure confidence intervals on the same plot and comment on the how the t interval compares to the other 2 intervals. You may use the `compare_CI` function provided below to produce the plot.

```
compare_CI <- function(ybar, QNORM, BOOT, TPROCEDURE,
                       col = c("#E41A1C","#377EB8","#4DAF4A")) {

  dt <- data.frame(type = c("qnorm", "bootstrap", "t"),
                   ybar = rep(ybar, 3),
                   low = c(QNORM[1], BOOT[1], TPROCEDURE[1]),
                   up = c(QNORM[2], BOOT[2], TPROCEDURE[2])
  )

  plot(dt$ybar, 1:nrow(dt), pch = 20, ylim = c(0, 5),
       xlim = range(pretty(c(dt$low, dt$up))),
       xlab = "Depth of ocean (m)", ylab = "Confidence Interval Type",
       las = 1, cex.axis = 0.8, cex = 3)

  abline(v = 37, lty = 2, col = "black", lwd = 2)
  segments(x0 = dt$low, x1 = dt$up,
           y0 = 1:nrow(dt), lend = 1,
           col = col, lwd = 4)

  legend("topleft",
         legend = c(eval(substitute( expression(paste(mu," = ",37)))),
                    sprintf("qnorm CI: [%.f, %.f]",QNORM[1], QNORM[2]),
                    sprintf("bootstrap CI: [%.f, %.f]",BOOT[1], BOOT[2]),
                    sprintf("t CI: [%.f, %.f]",TPROCEDURE[1], TPROCEDURE[2])),
         lty = c(1,1,1,1),
         col = c("black",col), lwd = 4)
}

# example of how to use the function:
compare_CI(ybar = 36, QNORM = c(25,40), BOOT = c(31, 38), TPROCEDURE = c(28, 40))
```