Assignment 4 - Central Limit Theorem, Confidence Intervals and Bootstrap. Due October 5, 11:59pm 2018

EPIB607 - Inferential Statistics[a]

[a]Fall 2018, McGill University

This version was compiled on October 3, 2018

In this assignment you will practice calculating confidence intervals. 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 5, 2018, 11:59pm. Both HTML and .Rmd files should be saved as 'IDnumber_LastName_FirstName_EPIB607_A4'.

Sampling distribution | Standard error | Normal distribution | Quantiles | Percentiles | Z-scores

Template

The `.Rmd` template for Assignment 4 is available here

1. Where do you buy?

Consumers can purchase nonprescription medications at food stores, mass merchandise stores such as Kmart and Wal-Mart, or pharmacies. About 45% of consumers make such purchases at pharmacies. What accounts for the popularity of pharmacies, which often charge higher prices?

A study examined consumers' perceptions of overall performance of the three types of stores, using a long questionnaire that had asked about such things as "neat and attractive store", "knowledgeable staff", and "assistance in choosing among various types of nonprescription medication". A performance score was based on 27 such questions. The subjects were 201 people chosen at random from the Indianapollis telephone directory. Here are the means ($\bar{y}$) and standard deviations (s) of the performances scores for the sample:

Store type	$\bar{y}$	s
Food stores	18.67	24.95
Mass merchandisers	32.38	33.37
Pharmacies	48.6	35.62

a. What population do you think the authors of the study want to draw conclusions about?
b. What population are you certain they can draw conclusions about?
c. Give a 95% confidence interval for the mean performance for each type of store.
d. Based on these confidence intervals, are you convinced that consumers think that pharmacies offer higher performance than the other types of stores?

2. Deer mice

Deer mice are small rodents native to North America. Their adult body lengths (excluding tail) are known to vary approximately Normally, with mean $\mu = 86$ millimeters (mm) and standard deviation $\sigma = 8$ mm. Deer mice are found in diverse habitats and exhibit different adaptations to their environment. A random sample of 14 deer mice in a rich forest habitat gives an average body length of $\bar{y} = 91.1$ mm. Assume that the standard deviation σ of all deer mice in this area is also 8 mm.

a. What is the standard deviation of the mean length $\bar{y}$?
b. What critical value do you need to use in order to compute a 95% confidence interval for the mean μ
c. Give a i) 90% confidence interval and ii) 95% confidence interval for the mean body length of all deer mice in the forest habitat.
d. Why does the 90% interval have a smaller margin of error?

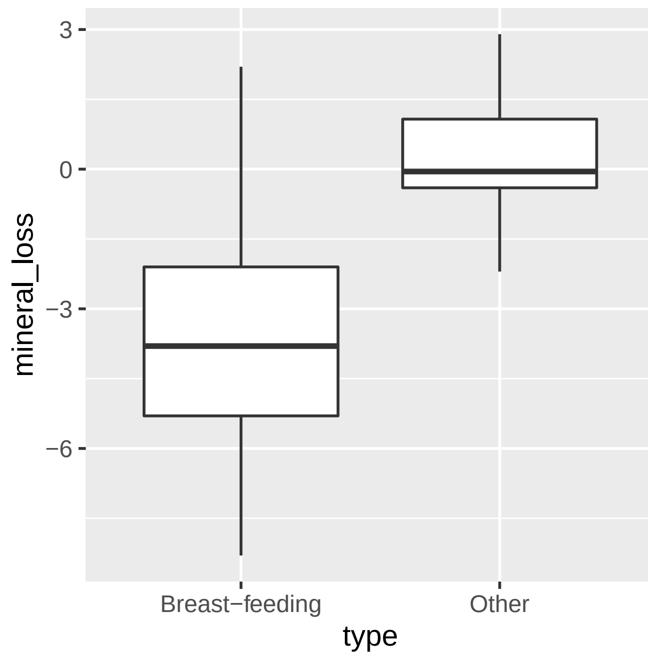
3. Does breast-feeding weaken bones?

Breast-feeding mothers secrete calcium into their milk. Some of the calcium may come from their bones, so mothers may lose bone mineral content. Researchers compared 47 breast-feeding women with 22 women of similar age who were neither pregnant nor lactating. They measured the percent change in the mineral content of the women's spines over three months. A negative value of `mineral_loss` indicates a loss in mineral content. The data can be read into `R` and a boxplot comparing the two groups is shown in the Figure below:

```r
boneloss <- read.csv("https://github.com/sahirbhatnagar/EPIB607/raw/master/data/boneloss.csv")
head(boneloss)
```

```
#      type mineral_loss
#   1 Other          2.4
#   2 Other          0.0
#   3 Other          0.9
#   4 Other         -0.2
#   5 Other          1.0
#   6 Other          1.7
```

```r
library(mosaic)
ggformula::gf_boxplot(mineral_loss ~ type, data = boneloss)
```



a. Give a 95% confidence interval for the mean mineral loss for each group.
b. Based on these confidence intervals, are you convinced that the data show distinctly greater bone mineral loss among the breast-feeding women?
c. (**BONUS**) Let $\bar{y}_1$ be the mean mineral loss for breast-feeding women and $\bar{y}_2$ be the mean mineral loss for the other group. Construct a 95% bootstrap confidence interval for the difference $\bar{y}_2 - \bar{y}_1$ based on $B = 10000$ bootstrap samples. Interpret this confidence interval and compare it to the ones obtained in b. *Hint*: split the data in two data.frames by `type`. For each data.frame, create B resamples and calculate the means for each. Plot the histogram of the difference in means, which gives you the bootstrap distribution for the difference. From this you can use the `quantile` function to calculate the 95% CI.

4. How deep is the ocean?

This question is based on the in-class Exercise on sampling distributions. Refer to the slides on Bootstrap confidence intervals for `R` code on how to calculate bootstrap confidence intervals. For your sample of $n = 20$ of depths of the ocean, calculate the

a. 95% Confidence interval using the ± formula
b. 95% Confidence interval using the `qnorm` function
c. 95% Confidence interval using $B = 10000$ bootstrap samples
d. Plot all three confidence intervals on the same plot and comment on the difference/similarities between the 3 intervals. You may use the `compare_CI` function provided below to produce the plot. This takes as input, the sample mean (`ybar`), and the CIs calculated from a,b,c in the form of a numeric vector of length 2 into the arguments `PM`, `QNORM` and `BOOT`, respectively.

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

  dt <- data.frame(type = c("plus_minus", "qnorm", "bootstrap"),
                   ybar = rep(ybar, 3),
                   low = c(PM[1], QNORM[1], BOOT[1]),
                   up = c(PM[2], QNORM[2], BOOT[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("plus/minus CI: [%.f, %.f]",PM[1], PM[2]),
                    sprintf("qnorm CI: [%.f, %.f]",QNORM[1], QNORM[2]),
                    sprintf("bootstrap CI: [%.f, %.f]",BOOT[1], BOOT[2])),
         lty = c(1, 1,1,1),
         col = c("black",col), lwd = 4)
}

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