Assignment 4 - Central Limit Theorem, Confidence Intervals and Bootstrap. Due October 15, 11:59pm 2021

EPIB607 - Inferential Statistics[a]

[a] Fall 2021, McGill University

This version was compiled on October 4, 2021

All questions are to be answered in an R Markdown document using the provided template and compiled to a pdf document. 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. Each question is worth 25 points. Label your graphs appropriately with proper titles and axis labels. Justify your answers. You may compile your reoport to pdf or to HTML. If you compile to HTML, then you must print the resulting HTML to pdf. Please submit the compiled pdf report to Crowdmark. You must also submit your code to Crowdmark. If you use the template, the code from your assignment will automatically appear at the end. Upload this code to Q5 in Crowdmark. You can upload a single pdf to Crowdmark, and then select the pages for a given question. See https://crowdmark.com/help/ for details.

Template

Use the template from the previous assignment.

1. (25 points) Concordance between PCR-based extraction-free saliva and nasopharyngeal swabs for SARS-CoV-2 testing - PART I

This question is based on the article *Concordance between PCR-based extraction-free saliva and nasopharyngeal swabs for SARS-CoV-2 testing*. The data used to reproduce the results is provided with the article and it provides Ct values for both test types (Nasopharyngeal and Saliva). Download the data, and use the following code to read it into R. Note that a Ct value of `undetected` implies that no virus was found in the sample. In the following R code, I specify `undetected` to be NA:

```r
library(readxl)
library(dplyr)
library(here)

# read symptomatic cohort data
dt_symp <- readxl::read_xlsx(
  here::here("Ct_values_for_matched_NPS_and_saliva_samples_(symptomatic_cohort).xlsx"),
  na = "undetected",
  col_names = c("ID","Nasopharyngeal","Saliva"),
  skip = 1,
  col_types = c("text", "numeric","numeric")
) %>%
  dplyr::mutate(cohort = "Symptomatic")

# read asymptomatic cohort data
dt_asymp <- readxl::read_xlsx(
  here::here("Ct_values_for_matched_NPS_and_saliva_samples_(asymptomatic_cohort).xlsx"),
  na = "undetected",
  col_names = c("ID","Nasopharyngeal","Saliva"),
  skip = 1,
  col_types = c("text", "numeric","numeric")
) %>%
  dplyr::mutate(cohort = "Asymptomatic")

# combine symptomatic and asymptomatic data together
dt <- dplyr::bind_rows(dt_symp, dt_asymp) %>%
  dplyr::mutate(cohort = factor(cohort))
```

a) (5 points) Reproduce Table 2. Hint: you can use the following command to create a new variable which indicates if the individual was SARS-CoV-2 negative or positive. To answer this question, you can simply show your code and it's output.

```r
ifelse(is.na(dt$Nasopharyngeal),"negative","positive")
```

b) (5 points) Reproduce the point estimates for Positive agreement and Negative agreement in Table 2. To answer this question, you can simply show your code and it's output.

c) (5 points) Can you determine which statistical procedure was used to calculate the confidence intervals for the point estimates in part b)? If yes, state the assumptions of the statistical procedure. If no, compute confidence intervals using a procedure of your choice and compare them with the ones given in the paper. State your assumptions.

d) (5 points) Consider the 3 figures shown in Figure 1. Comment on the dataset format that would have been used to create the three figures (e.g. tidy format, wide format).

e) (5 points) Create a graphic which would support the claim that *The difference in mean Ct values (0.132) was not statistically significant (p=0.860)*. You can either recreate a graph shown in the paper, or create your own. Justify why your figure supports the claim. Note: you **are not** being asked to perform any statistical test.

2. (25 points) Concordance between PCR-based extraction-free saliva and nasopharyngeal swabs for SARS-CoV-2 testing - PART II

This question is based on the article *Concordance between PCR-based extraction-free saliva and nasopharyngeal swabs for SARS-CoV-2 testing*. The code to read in the data is shown in the previous question. For the following questions, only use the symptomatic cohort.

a. (5 points) Give a 95% confidence interval for the mean Ct value for each test type (Nasopharyngeal and Saliva). State your assumptions.

b. (5 points) Based on these confidence intervals, are you convinced that the data show there is no difference in mean Ct values between both test types? Explain.

c. (5 points) Construct a 95% confidence interval for the mean difference in Ct values between both test types using a canned function. Interpret this confidence interval. State the assumptions being made by this function.

d. (5 points) Construct a 95% bootstrap confidence interval for the mean difference in Ct values between both test types. Compare it to the one obtained in part c. State your assumptions.

e. (5 points) Compute a standard error for the mean difference in Ct values between both test types. Was the standard error that you computed used in any of the confidence intervals you constructed in parts a,c or d? Explain.

3. (25 points) Concordance between PCR-based extraction-free saliva and nasopharyngeal swabs for SARS-CoV-2 testing - PART III

This question is based on the article *Concordance between PCR-based extraction-free saliva and nasopharyngeal swabs for SARS-CoV-2 testing*. The code to read in the data is shown in the previous question.

a. (5 points) For the asymptomatic cohort, calculate the sensitivity and specificity of the saliva test. Show your work.

b. (7.5 points) The reviewer reports are provided with the article. Specifically, Reviewer 2 has an issue with the very wide confidence interval for sensitivity of the saliva test. Without performing any calculations, do you agree or disagree with their statement. Explain.

c. (7.5 points) Are you able to calculate a 95% bootstrap confidence interval for the sensitivity of the saliva test? If yes, compute it and compare it to the one given in the paper. If not, explain why.

d. (5 points) Do you think the 95% confidence interval provided for the specificity is correct? Explain.

4. (25 points) How deep is the ocean?

This question is based on the in-class exercise on sampling distributions using the depths of the ocean example.

a. (5 points) For your sample of $n = 5$ of depths of the ocean, calculate the 95% Confidence interval using the ± formula, the `qnorm` function, and using $B = 10000$ bootstrap samples.

b. (5 points) 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.

c. (10 points) Repeat parts a and b using your sample of size $n = 20$. Comment on the difference/similarities between the $n = 5$ and $n = 20$ graph.

d. (5 points) In the in-class exercise, there were a total of approximately $N = 60$ students who participated. In your own words, briefly explain the difference between the $N = 60$, and the $n = 5$ or $n = 20$. What effect did these different 'n' have on the sampling distribution of the sample means?

```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))
```