Assignment 7 - Sample Size, Proportions, Rates, Linear Regression. Due November 19, 11:59pm 2021

EPIB607 - Inferential Statistics[a]

[a]Fall 2021, McGill University

This version was compiled on November 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) REGEN-COV Antibody Combination and Outcomes in Outpatients with Covid-19

This question is based on the paper *REGEN-COV Antibody Combination and Outcomes in Outpatients with Covid-19* by Weinreich et al. 2021 published in the New England Journal of Medicine (available on myCourses).

a) (10 points) Use a simulation based approach (as shown in class) to reproduce the sample size calculation for the average change from baseline in viral shedding from day 1 to day 22 (details can be found in Section 11.2 of the study protocol also available on myCourses). Specifically focus on the sample size of 20 patients per arm in phase 1. Hint: you are being asked to show (via simulation) that the study is indeed powered at the stated value to detect the stated effect.

b) (10 points) The sample size section of the protocol continues further:

> Assuming a 10% dropout rate and standard deviation of 2.1 log10 copies/mL (Cao, 2020), a sample size of 50 patients per arm in phase 2 will have at least 80% power to detect a difference of 1.25 log10 copies/mL.

Using simulations, reproduce this power calculation.

c) (5 points) Finally, the protocol says:

> If a standard deviation of 3.8 log10 copies/mL is assumed (Wang, 2020c), the detectable difference would be 2.27 log10 copies/mL

Use the `pwr` R package (https://cran.r-project.org/web/packages/pwr/index.html) to reproduce this power calculation.

2. (25 points) Simulation study for confidence intervals of proportions

The goal of this simulation study is to estimate the coverage probabilities of confidence intervals for the binomial proportion. In class we saw at least three methods to calculate the confidence interval for a proportion: 1) normal approximation, 2) the exact method (Clopper-Pearson) and 3) the plus 4 method.

a) (10 points) Simulate 100 trials from a Binomial(n = 10, $\pi = 0.1$) distribution using the `stats::rbinom` function. For each of these trials, calculate the 95% confidence interval for the proportion using each of the three methods mentioned above. For each of the three methods: plot the confidence intervals and color each of them by whether they covered the truth or not. Hint: see the R code for the slides on one sample rates.

b) (5 points) For each of the three methods, calculate the coverage probability, i.e., the percentage of intervals which contain the true population proportion π. Describe your findings and comment on the differences between methods in terms of coverage probability.

c) (10 points) Repeat the simulation study in a), but this time, with different combinations of n and π. Visualize the coverage probabilities as a function of n and π for each of the three methods. Describe your findings and comment on the differences between methods in terms of coverage probabilities as a function of n and π. What happens to the coverage probabilities when the expected number of events increases? Explain.

3. (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:

```
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) (8 points) For the symptomatic cohort, was there a difference in mean Ct values? Use an appropriate regression model to answer this question. Write the regression equation in terms of population parameters and define all parameters in your model. What is the parameter of interest?

b) (5 points) Estimate the regression parameters using the data provided. Report the estimated coefficient for the parameter of interest and interpret it in the context of the problem.

c) (5 points) Reproduce the p-value for the parameter of interest from the regression output and interpret it. What assumptions were used in calculating the p-value. (Note: you are being asked to show how the p-value was calculated. Do not simply copy the value from the regression output.)

d) (7 points) Use a permutation test to calculate the p-value for the parameter of interest and compare it with the one obtained in part c). Briefly discuss this comparison.

4. (25 points, 5 each) Physical activity in NHANES

This problem uses data from the National Health and Nutrition Examination Survey (NHANES), a survey conducted annually by the US Centers for Disease Control (CDC). The dataset is available from the `NHANES` package.

Regular physical activity is important for maintaining a healthy weight, boosting mood, and reducing risk for diabetes, heart attack, and stroke. In this problem, you will be exploring the relationship between weight (`Weight`) and physical activity (`PhysActive`) using the NHANES data. Weight is measured in kilograms. The variable `PhysActive` is coded `Yes` if the participant does moderate or vigorous-intensity sports, fitness, or recreational activities, and `No` if otherwise. The objective is to compare weight between physically active and those who are not.

a) Explore the data.

   i. Identify how many individuals are physically active.

   ii. Create a plot that shows the association between weight and physical activity. Describe what you see.

b) Provide an appropriate regression model for the stated objective and state the parameter of interest. Give the regression equation in terms of population parameters and be sure to define each of the parameters in your model.

c) Fit a linear regression model to estimate the regression parameters. Report the estimated coefficients from the model and interpret each of them in the context of the problem.

d) Report a 95% confidence interval for the parameter of interest and interpret the interval in the context of the problem. Based on the interval, is there sufficient evidence at $\alpha = 0.05$ to reject the null hypothesis of no association between weight and physical activity? State the assumptions used for calculating the 95% confidence interval.

e) Provide a 95% bootstrap confidence interval for the parameter of interest and compare it to the one in part d). Briefly discuss the comparison.