Assignment 1 - Exploring Data. Due September 25, 11:59pm 2019

EPIB607 - Inferential Statistics[a]

[a] Fall 2020, McGill University

This version was compiled on September 11, 2020

All questions are to be answered 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. Each question is worth 25 points. Label your graphs appropriately with proper titles and axis labels. Please submit both the compiled HTML report and the source file (.Rmd) to myCourses by September 22, 2019, 11:59pm. Both HTML and .Rmd files should be saved as 'IDnumber_LastName_FirstName_EPIB607_A2'.

Histograms | Bar plots | Line plots | ggformula package | mosaic package

Template

Please use the `.Rmd` template for Assignment 1 is available on myCourses.

1. (25 points) Immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2

This questions refers to the Lancet paper *Safety and immunogenicity of the ChAdOx1 nCoV-19 vaccine against SARS-CoV-2:a preliminary report of a phase 1/2, single-blind, randomised controlled trial* by Folegatti et. al (2020) and available in myCourses.

a) (2 points) Consider Figure 3 Panel B: What visual cues (or aesthetics) are being used? Briefly describe the main takeaways from the entire Figure 3.
b) (3 points) Do you think Figure 3 is a good graphic in terms of conveying its message clearly? Is there anything you would have done differently? Explain.
c) (2 points) Consider the data introduced in class which contains immunity levels (Immunoglobulin G (IgG)) from the convalescent group and the vaccine groups post 28 days. Note that the IgG levels in the dataset below are given on the log10 scale. Calculate the median IgG levels (ELISA units) on the log10 scale for each group.

```
path <-
"http://www.biostat.mcgill.ca/hanley/statbook/immunogenicityChAdOx1.nCoV-19vaccine.txt"
ds <- read.table(path)
head(ds)
```

```
#   RefIndexCategory IgGResponse.log10.ElisaUnits
# 1     Convalescent                         2.56
# 2     Convalescent                         2.74
# 3     Convalescent                         2.79
# 4     Convalescent                         3.32
# 5     Convalescent                         3.15
# 6     Convalescent                         2.35
```

```r
str(ds)
```

```
#  'data.frame':    307 obs. of  2 variables:
#   $ RefIndexCategory            : Factor w/ 2 levels "Convalescent",..: 1 1 1 1 1 1 1 1
#   $ IgGResponse.log10.ElisaUnits: num  2.56 2.74 2.79 3.32 3.15 2.35 2.72 2.95 2.42 2.64
```

d) (4 points) From the medians alone, is there enough evidence to conclude that the median IgG levels in the convalescent group are higher than the median IgG levels in the vaccine group (post 28 days)? Explain.

e) (7 points) Use the Boostrap to asses if there is enough evidence to suggest that the median IgG levels in the convalescent group are higher than the median IgG levels in the vaccine group (post 28 days). *Hint: resample the data with replacement separately in each group B=1000 times. For each of the B datasets, calculate the median IgG level and take the difference in medians between the two groups. Plot the differences in a histogram and calculate the 2.5 and 97.5 percentiles.*

f) (7 points) The dataset, shown below and available on myCourses, was extracted (approximately) from Figure 3 Panel A for the ChAdOx1 nCoV-19 (prime) group only. The `time` column represents the days since vaccination, and `igg_response` are the IgG levels on the original scale. Create an appropriate figure which shows the immunity levels as a function of time. You are free to choose the plot type; the choice of plot should be guided by the message you are trying to convey. Be sure to label your axes, show units, include a title and choose an appropriate color palette. Briefly interpret the plot.

```r
DT <- read.csv("prime_igg_response.csv")
```

```r
head(DT)
```

```
#    time igg_response
#  1    0    930.37376
#  2    0    267.80142
#  3    0    241.40290
#  4    0    170.80787
#  5    0     79.79795
#  6    0     67.12348
```

2. (25 points) COVID-19 Cases Comparison Between Counties With and Without Stay-at-Home Orders

This question is based on the JAMA Network Open article *Comparison of Estimated Rates of Coronavirus Disease 2019 (COVID-19) in Border Counties in Iowa Without a Stay-at-Home Order and Border Counties in Illinois With a Stay-at-Home Order* by Lyu and Wehby (2020) and available in myCourses. The county and state level cumulative incidence of cases data is provided in the code below. Note: you need to install the `covdata` package (which is not on CRAN) using `remotes::install_github("kjhealy/covdata")`.

```r
# remotes::install_github("kjhealy/covdata")
library(covdata)
library(dplyr); library(tidyr); library(ggplot2); library(readr)
# get population data from https://covid19.census.gov/datasets/
f <- "https://opendata.arcgis.com/datasets/21843f238cbb46b08615fc53e19e0daf_1.csv"
pop_county <- read_csv(file = f) %>%
  dplyr::rename(fips = GEOID, population = B01001_001E, state = State) %>%
  dplyr::select(state, fips, population)

county_level <- nytcovcounty %>%
  dplyr::left_join(pop_county, by = c("state","fips")) %>%
  dplyr::mutate(cases.per.10k = cases/population * 1e4) %>%
  dplyr::filter(state %in% c("Iowa","Illinois")) %>%
  dplyr::group_by(county)

pop_state <- pop_county %>%
  dplyr::group_by(state) %>%
  dplyr::summarise(population = sum(population, na.rm = TRUE))

state_level <- county_level %>%
  dplyr::group_by(state, date) %>%
  dplyr::filter(date >= "2020-03-15") %>%
  dplyr::summarise(cases = sum(cases)) %>%
  dplyr::left_join(pop_state, by = "state") %>%
  dplyr::mutate(cases.per.10k = cases / population * 1e4, state = factor(state),
                time = as.numeric(date - min(date)) + 1)
```

```r
head(county_level)
```

```
#  # A tibble: 6 x 8
#  # Groups:   county [1]
#    date       county state    fips  cases deaths population cases.per.10k
#    <date>     <chr>  <chr>    <chr> <dbl>  <dbl>      <dbl>         <dbl>
#  1 2020-01-24 Cook   Illinois 17031     1      0    5223719       0.00191
#  2 2020-01-25 Cook   Illinois 17031     1      0    5223719       0.00191
#  3 2020-01-26 Cook   Illinois 17031     1      0    5223719       0.00191
#  4 2020-01-27 Cook   Illinois 17031     1      0    5223719       0.00191
#  5 2020-01-28 Cook   Illinois 17031     1      0    5223719        0.00191
#  6 2020-01-29 Cook   Illinois 17031     1      0    5223719        0.00191
```

```
head(state_level)
```

```
#  # A tibble: 6 x 6
#  # Groups:   state [1]
#    state    date       cases population cases.per.10k  time
#    <fct>    <date>     <dbl>      <dbl>         <dbl> <dbl>
#  1 Illinois 2020-03-15    94   12821497        0.0733     1
#  2 Illinois 2020-03-16   104   12821497        0.0811     2
#  3 Illinois 2020-03-17   159   12821497        0.124      3
#  4 Illinois 2020-03-18   286   12821497        0.223      4
#  5 Illinois 2020-03-19   420   12821497        0.328      5
#  6 Illinois 2020-03-20   583   12821497        0.455      6
```

a) (6 points) Using the county level dataset provided, reproduce Figure 1 of the paper. Does your Figure agree with theirs? Would county level curves have been more appropriate to show instead of the state totals?

b) (5 points) Plot the cumulative incidence curves per 10000 people from March 21 until the most recent day for which you have data, for each of the counties used in the paper. Interpret the plot and discuss if the county level plots still agree with the overall conclusion of the paper.

c) (4 points) Case counts are inherently tied to testing capacity. Death from COVID19 doesn't have this issue, although there are other biases such as misclassification and under reporting. Plot the same graph as in part (b) but for deaths and interpret the plot.

d) (10 points) Illinois (Democrat-controlled legislature) is surrounded by states with Republican-controlled legislatures (Iowa, Missouri, Kentucky, Indianna, Wisconsin). Do the data suggest there is a correlation between COVID-19 cases (or deaths) and which party has legislative control? Explain and justify using summary statistics and/or figures.

3. (25 points) Age-structures of Populations, then and now

The 1911 census of Ireland was taken on April 2nd 1911 and was released to the public in 1961. Follow this link for further details on the census. James Hanley (JH) has scrapped the data for Dublin, collected the age-frequency distribtion by gender and provided you with a three column .csv file on myCourses called `age_sex_frequencies_ireland.csv` which looks like this:

```r
cens <- read.csv("age_sex_frequencies_ireland.csv")
```

```r
head(cens)
```

```
#   Gender Age Freq
#  1   Male   0 5332
#  2   Male   1 4570
#  3   Male   2 4979
#  4   Male   3 4789
#  5   Male   4 4884
#  6   Male   5 4787
```

The `Age` column represents the age in 1911. The `Freq` column gives the frequency of the number of people for a given age and `Gender`. Note that `Age` is an interval; for example, `Age=0` actually represents individuals who are between the ages of 0 and 1, `Age=1` are individuals between ages 1 and 2, and so on.

a) (6 points) What was the earliest year of birth for (i) males and (ii) females ?
b) (8 points) Create a suitable visualization of this data and then comment on any patterns you see and give reasons for these patterns. Your choice should leverage all the information provided in the data and be influenced by the message that you are trying to convey. Be sure to include an informative title and figure caption.
c) (8 points) Calculate the mean age, the standard deviation (SD), and the quartiles: $Q_{25}, Q_{50}(median), Q_{75}$ separately for males and females.
d) (3 points) The original census cards have been scanned are available online. This one in particular is quite famous. Why?

4. (25 points) Flint Blood Lead Levels

Lead in the environment is persistent, bio-accumulative, and toxic. Chronic exposure to lead in children is associated with many negative health outcomes even when the Blood Lead Levels (BLLs) are measured as low as 1.0-10.0 μg/dL. An analysis of childhood exposure to lead is described in the article *Blood Lead Levels of Children in Flint, Michigan: 2006-2016* by Gomez et al. (2018) available on myCourses.

a) (2 points) As presented in, Figure 3, is BLL on a continuous or discrete scale? Is Frequency on a continuous or discrete scale?

b) (3 points) Briefly comment on a strength of Figure 3. i.e.: what information jumps out at you when you look at this figure?

c) (3 points) Briefly comment on a weakness of Figure 3. i.e.: what information is presented in the figure, but is more difficult to interpret/see?

d) (17 points: 5 for data entry, 6 for visualization, 3 for advantage, and 3 for disadvantage) Figure 3 data in tabular form is show below. Create a `data.frame` of the data and a new visualization using the data. Explain why your data visualization is better and/or worse than Figure 3. The visualization does not need to be radically different than Figure 3, but it can be if you are feeling especially rad today. The aim is to make a visualization that is *not* a bar chart and to explain at least one advantage and one disadvantage of your new and exciting visualization with respect to Figure 3. You may facet the visualization as long as you justify your choice. In the data shown below, the `bll_cat` column is the category of BLL and `cat_freq` is the frequency percentage of the particular BLL category in a given year. All values are taken from Figure 3 and the values for `cat_freq` are approximate. *Note that we have not provided you with this dataset. Part of this exercise is for you to practice how to extract data from a graph and enter it in R.*

```r
head(bll_df, n= 4)
```

```
#    year bll_cat cat_freq
#  1 2012    <0.5      2.0
#  2 2013    <0.5      4.5
#  3 2014    <0.5      6.0
#  4 2015    <0.5      4.5
```

```r
str(bll_df)
```

```
#  'data.frame':    35 obs. of  3 variables:
#   $ year    : int  2012 2013 2014 2015 2016 2012 2013 2014 2015 2016 ...
#   $ bll_cat : Factor w/ 7 levels "<0.5",">5.0",..: 1 1 1 1 1 3 3 3 3 3 ...
#   $ cat_freq: num  2 4.5 6 4.5 6 28.5 37 35 35 44.5 ...
```

```r
levels(bll_df$bll_cat)
```

```
#  [1] "<0.5"    ">5.0"    "0.5-1.0" "1.1-2.0" "2.1-3.0" "3.1-4.0" "4.1-5.0"
```