STAT505 Data Analysis Overview

One interesting characteristic of Severe Acute Respiratory Syndrome coronavirus 2 (SARS-CoV-2), the virus that causes COVID-19, is the prevalence of asymptomatic cases and the ability of asymptomatic carriers to infect others. While there are many public health implications of asymptomatic spread, this question will explore clinical and immunological measurements of symptomatic and asymptomatic patients. Specifically of interest will be antibodies which are proteins in the blood that develop to neutralize, in this case, the SARS-CoV-2 virus. Furthermore, the presence of antibodies is commonly linked with immunity to future infections.

An recent article published in Nature Medicine titled *Clinical and immunological assessment of asymptomatic SARS-CoV-2 infections* contains a comparative study of symptomatic and asymptomatic patients. The paper freely available at the **following link**.

NOTE: A question motivated by this dataset was on the Fall 2020 comprehensive exams.

```
library(tidyverse)
```

1. Download Data

```
## -- Attaching packages --------------------- tidyverse 1.3.0 --
```

```
## v ggplot2 3.3.1     v purrr   0.3.4
## v tibble  3.0.1     v dplyr   1.0.0
## v tidyr   1.1.0     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------ tidyverse_conflicts() --
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
Covid_data <- read_csv("http://math.montana.edu/ahoegh/Data/Covid_3a.csv")
```

```
## Parsed with column specification:
## cols(
##   `Patient ID` = col_double(),
##   `IgG S/CO` = col_double(),
##   `IgG (+/-)` = col_character(),
##   `IgM S/CO` = col_double(),
##   `IgM  (+/-)` = col_character(),
##   Group = col_character()
## )
```

```
Covid_data
```

2. View a few rows of the dataset.

```
## # A tibble: 74 x 6
##    `Patient ID` `IgG S/CO` `IgG (+/-)` `IgM S/CO` `IgM  (+/-)` Group
```

```
##           <dbl>      <dbl> <chr>           <dbl> <chr>        <chr>
##  1            1     12.3   +               3.83  +            Asymptomatic
##  2            2      3.14  +               1.14  +            Asymptomatic
##  3            3      0.304 -               0.173 -            Asymptomatic
##  4            4      1.76  +               0.138 -            Asymptomatic
##  5            5      0.532 -               0.352 -            Asymptomatic
##  6            6      1.74  +               2.46  +            Asymptomatic
##  7            7      0.531 -               4.23  +            Asymptomatic
##  8            8     10.7   +               2.33  +            Asymptomatic
##  9            9     13.4   +               4.58  +            Asymptomatic
## 10           10      1.03  +               1.88  +            Asymptomatic
## # ... with 64 more rows
```

3. Research Question formulation *Note for writing in this class and the entire MS/PhD Stats program, I'd highly recommend Writing Science by Schimel.*

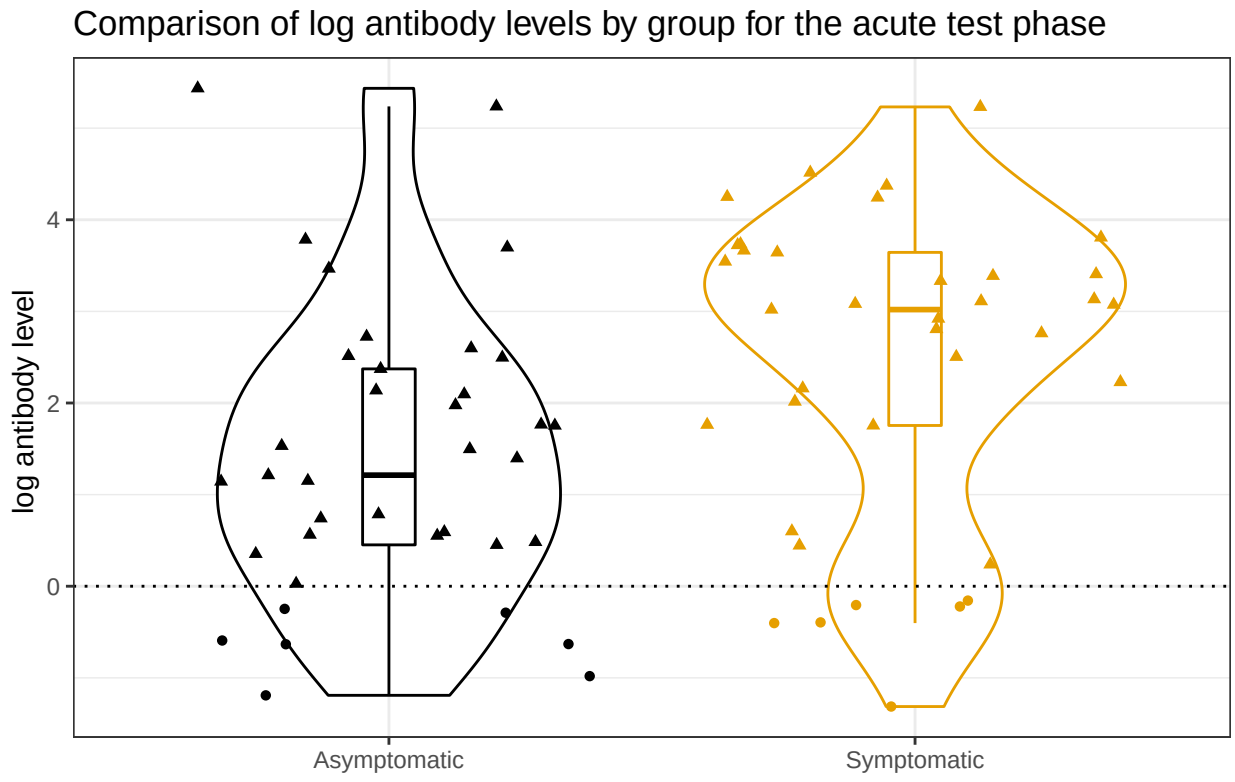
Often this step will be done collaboratively with other researchers or scientists. For now, lets assume the research question is "are differences in the IgG antibodies between the symptomatic or asymptomatic group?"

Q: how do we feel about this question? Is it specific enough and answerable with the data?

4. Data Visualization Next, we explore the raw data using `ggplot2`.

```
cb_pal <- c("#000000", "#E69F00", "#56B4E9", "#009E73",
            "#F0E442", "#0072B2", "#D55E00", "#CC79A7") # colorblind friendly pallete

Covid_data  %>% mutate(log_IgG = log(`IgG S/CO`)) %>%
  ggplot(aes(y=log_IgG, x = Group, color = Group)) +
  geom_violin(width = .8) +
  geom_boxplot(width = .1, outlier.shape = NA) +
  theme_bw() +
  theme(legend.position = 'none')  +  scale_colour_manual(values=cb_pal) +
  geom_jitter(aes(shape = `IgG (+/-)`)) +
  ylab('log antibody level ') +
  xlab('') +
  ggtitle('Comparison of log antibody levels by group for the acute test phase') +
  geom_hline(yintercept = log(1), linetype = 3) +
  labs(caption = "Dashed line is the threshold for seropositive results.")
```



5. Refined Research Question Generally, the discussions that I hear are focused on whether patients are immune. So I'm going to focus on the proportion of seropositive patients by group.

6. Model Specification For binary data, a common approach is to use logistic regression. With logistic regression,

$$y_i \sim Bernoulli(p_i) \tag{1}$$
$$logit(p_i) = X_i\beta, \tag{2}$$

where y_i is a binary variable for whether the i^{th} observation is a seropositive, p_i is the probability that the i^{th} observation is a success,

$$X_i = \begin{bmatrix} I(Group[1] = Asymptomatic) & I(Group[1] = Symptomatic) \\ I(Group[2] = Asymptomatic) & I(Group[2] = Symptomatic) \\ I(Group[n] = Asymptomatic) & I(Group[n] = Asymptomatic) \end{bmatrix}$$

is a $n \times 2$ matrix of covariates, where $I(Group[i] = Asymptotic)$ is an indicator function for whether i^{th} patient is in the asymptotic group, and

$$\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix},$$

is an $n \times 1$ matrix, or vector, of covariates. $logit^{-1}(\beta_0)$ is the probability of a seropositive individual in the Asymptomatic group and $logit^{-1}(\beta_1)$ is the probability of a seropositive individual in the Symptomatic group. *Note this is the cell means model. This can also be formulated as a reference case model.*

7. Model Fit We fit this model using two different R functions. The first approach uses the `glm()` function.

```
library(arm)
Covid_data <- Covid_data %>% mutate(seropositive = `IgG (+/-)` == '+' )
glm_fit <- glm(seropositive ~ Group - 1, data = Covid_data, family = binomial(link = "logit"))
display(glm_fit)
```

```
## glm(formula = seropositive ~ Group - 1, family = binomial(link = "logit"),
##     data = Covid_data)
##                   coef.est coef.se
## GroupAsymptomatic 1.46     0.42
## GroupSymptomatic  1.64     0.45
## ---
##   n = 74, k = 2
##   residual deviance = 68.7, null deviance = 102.6 (difference = 33.9)
```

A second approach, with is featured prominently in the textbook, uses `stan_glm()`.

```
library(rstanarm)
glm_stanfit <- stan_glm(seropositive ~ Group - 1, data = Covid_data, family = binomial(link = "logit"),
print(glm_stanfit, digits = 2)
```

```
## stan_glm
##  family:       binomial [logit]
##  formula:      seropositive ~ Group - 1
##  observations: 74
##  predictors:   2
## ------
##                   Median MAD_SD
## GroupAsymptomatic 1.47   0.43
## GroupSymptomatic  1.68   0.44
##
## ------
## * For help interpreting the printed output see ?print.stanreg
## * For info on the priors used see ?prior_summary.stanreg
```

We see that the results are very similar (more later) for both functions and, furthermore, there is not a substantial difference (more later here too) between the two groups.