**Statistics with epidemiology of COVID-19**

Maria Shumskaya, Shakira Benjamin, Matthew Niepielko, Nicholas Lorusso. Kean University, NJ, USA.

**Student handout**

**Introduction**

The coronavirus disease 2019 (COVID-19) pandemic is a global challenge caused by a rapid emergence of the SARS-CoV-2 virus. Over the course of 2020, the numbers of infections reached 144,220,516 cases and 3,065,612 deaths worldwide.The most current information for cases and deaths can be found at at <https://www.worldometers.info/coronavirus/>. According to the United States Centers for Disease Control and Prevention (CDC) as of April 21, 2021 there are at least 31,602,676 reported cases and 565,613 total deaths in the United States, with 983,875 reported cases and 283,000 total deaths in the state of New Jersey alone. You can find the most recent data at <https://covid.cdc.gov/covid-data-tracker/#cases_totalcases>.

Prevention of infection, timely identification of infected individuals, and effective treatment are important in managing COVID-19 and easing the effects of the pandemic. The need for the most up to date epidemiological information has encouraged rapid action by both private and public entities to develop cost-effective testing devices and competent laboratory techniques for diagnosing SARS-CoV-2 infection, as well as recent development of several vaccines. ModernaTX, Inc and Pfizter-BioTNech have engineered mRNA vaccines (mRNA-1273 and BNT162B2) that are currently being administered in US and worldwide. Additionally, Janssen Pharmaceuticals Companies of Johnson & Johnson developed a viral vector vaccine, JNJ-78436735.

Testing for the SARS-CoV-2 viral RNA (PCR-based) recognizes infected individuals during the acute phase of an infection. Serological tests identify individuals who have established antibodies, Immunoglobulin G and M (IgG and IgM), against the virus. Combination of these two testing methods enhances the ability to carry out contact tracing to monitor the spread of the viral infection over time while still tracking active infections.

According to statistics produced by the CDC, several factors are important in predicting the likeliness of the infection and severity of the disease development such as age, gender, and density of a population. SARS-CoV-2 infection has been observed to be more prevalent in males compared to females in the United States. In addition to observations in the United States, the majority of cases reported globally contain an observable difference between male and female positive cases and hospitalization rates. Data from China show that 54.3–57.3% of hospitalized patients and 61.1% of ICU patients were male. Additionally, in Korea in-hospital deaths were 61.1% males. In Europe, studies observed that a 62% higher admission of males to inpatient care compared with females; and Italy’s fatality rate by March 2020 was reported to be 10.6% for males and 6% for females. Regional population density has been also observed to have an impact on infection rates with areas with higher population density having more reported positive cases and deaths.

**References and further reading**

Carter, L. J. et al, Assay Techniques and Test Development for COVID-19 Diagnosis. ACS central science 2020, 6 (5), 591-605.

Groban, L. et al., Is Sex a Determinant of COVID-19 Infection? Truth or Myth? Current Hypertension Reports 2020, 22 (9), 62.

Kadi N., Khelfaoui, M. Population density, a factor in the spread of COVID-19 in Algeria: statistic study. Bull Natl Res Cent. 2020, 44(1):138.

**Assignment**

In this exercise, we will analyze an original dataset prepared by Genesis Laboratory Management, a COVID-testing laboratory located in Monmouth County, NJ, USA. The laboratory collects patient specimens from the whole New Jersey, from all counties though more samples were received from Monmouth and surrounding counties. The samples were tested for SARS-CoV-2 RNA sequence presence using PCR and the results were recorded in the database. The data you will be working with cover the period from March to December 2020.

We will use RStudio, and, among others, R packages: maps, mapdata, ggplot2, gifski and dplyr.

The dataset used in this module is an adapted version from the original data from Genesis Laboratory Management. All personal identifying information was deleted from the dataset you will receive. You will be provided a CSV (comma delimited) file with individual test result entries as rows, and information on patient county, age, sex, and test results as columns. The data set also contains the central geographic location (longitude and latitude) for each county and the rank of the county’s population density, 1 most dense and 21 least dense.

**Preliminary questions:** You will be working with the original data collected for the state of New Jersey, USA.

1. Do you think the number of positive cases for COVID-19 has changed from March to December of 2020 in each county?
2. Do you think it will depend on the number of tests conducted, or the population of the county?
3. Do you think you will find a difference in the number of positive tests between males and females?
4. Was the virus spreading from one location within the state or from multiple location at the same time?

Formulate and record a hypothesis.

We will test our hypotheses by visualizing data using a choropleth map, which is a type of [thematic map](https://en.wikipedia.org/wiki/Thematic_map) in which a set of pre-defined areas is colored or patterned in proportion to a statistical variable. More simply, we will be referring to these maps as heat maps. We will also perform a variation of a t-test to compare male and female populations.

**PART 1. Heat maps and visualization of the number of positive tests for each county**

First, you will familiarize yourself with maps and mapdata R packages, which will help you to create beautiful maps using ggplot2 package. You can read more about maps here: <https://eriqande.github.io/rep-res-web/lectures/making-maps-with-R.html>, excerpts below:

***The main players:***

* The maps package contains a lot of data encoding outlines of continents, countries, states, and counties that have been built up in R over years.
* The mapdata package contains a few more, higher-resolution outlines for geographic data.
* ggplot2 operates on data frames. Therefore, we need some way to translate the maps data into a data frame format the ggplot can use.
* ggsave allows for the maps to be saved on your computer as a high resolution image

***Maps in the* maps *package***

* The maps package provides lots of different map outlines and points for cities, etc. Some examples: usa, nz, state, world, etc.

***Making data frames from map outlines***

* ggplot2 provides the map\_data() function. Think of it as a function that turns a series of points along an outline into a data frame of those points.
* Syntax: map\_data("name") where “name” is a quoted string of the name of a map in the maps or mapdata package.

***The structure of the data frames***

* long is longitude. Things to the west of the prime meridian are negative.
* lat is latitude.
* order. This shows in which order ggplot should “connect the dots”
* region and subregion tell what region or subregion a set of points surrounds.
* group. This is *very important*! ggplot2’s functions can take a group argument that controls (amongst other things) whether adjacent points should be connected by lines. If they are in the same group, then they get connected, but if they are in different groups then they don’t. Essentially, having to points in different groups means that ggplot “lifts the pen” when going between them.

***Plot the USA map***

* Maps in this format can be plotted with the polygon geom. i.e. using geom\_polygon().
* geom\_polygon() draws lines between points and “closes them up” (i.e. draws a line from the last point back to the first point)
* You have to map the group aesthetic to the group column (aesthetic means something you can see, e.g. position on axis or color)
* x = long and y = lat are the other aesthetics.

A map of USA in your script will be drawn like this:

usa = map\_data("usa")

USA\_base = ggplot(data = usa, mapping = aes(x = long, y = lat, group = group)) +

coord\_fixed(1.3) +

geom\_polygon(color = "black", fill = "white")

USA\_base

coord\_fixed()is very important when drawing maps. It fixes the relationship between one unit in the *y* direction and one unit in the *x* direction. Then, even if you change the outer dimensions of the plot (i.e. by changing the window size or the size of the pdf file) the *aspect ratio* remains unchanged. In the above case, every *y* unit is 1.3 times longer than an *x* unit, then the plot comes out looking good. A different value might be needed closer to the poles.

**PART 2. Wilcoxon–Mann–Whitney test**

You will compare the percent of positive results for SARS-CoV-2 between male and female populations. In order to do it, you will use a Wilcoxon–Mann–Whitney test. This nonparametric test that will return a p-value for the null hypothesis that for randomly selected values X and Y from two populations, the probability of X being greater than Y is equal to the probability of Y being greater than X. If p-value is less than 0.05, then the probability that the null hypothesis is true is 5% and we can accept the alternate hypothesis – that the male and female populations are significantly different in the number of positive test results.

To run such a test, you will use wilcox.test on two subsets, male and female, that you will create out of the covid19Data.

**PROCEDURE**

Open TeachingModule\_KeanU\_COVID19NJ\_students.R in R studio. Import the dataset covid19Data.csv. Follow the steps in the script. Be sure to save your workspace when you are done by clicking on the environment tab on the right side and clicking the blue disc icon. Save the workspace environment as NJ\_Covid.RData. This will allow you to return to your data later to continue with additional analyses.

**Please answer the questions below and include any requested figures:**

1. What kind of information is stored in columns subregion, countyPop, MONTH, TEST\_NUMERICAL?
2. Click on Zoom in your right bottom frame of RStudio. Right click on image and save it as PNG or JPEG, then insert it below. Alternatively, you can use Export – Save as Image – then use PNG, TIFF or JPEG to save and then insert it to this file. Adjust the image dimensions as necessary to capture key image features.
3. Open the working directory location on your computer and paste the image that was automatically saved by the ggsave function. What do different saturations of blue colors represent? What can you say about the number of positive tests in New Jersey in 2020 according to this map? What county appears to have the highest number of positive cases per person?
4. Insert the March map here. What can you say about the number of positive tests in New Jersey in March 2020 specifically according to this map? How is this different from the previous map in Question 3?
5. Save the map for June and insert it here.
6. Save the map for December and insert it here.
7. Compare the maps for March, June and December. What trends do you see? What factors do you think affect the spread of the virus across the counties?
8. What number did you get and what do you think length function does?
9. What number did you get and what do you think sum function does?
10. Save the bar plot as image and insert it here.
11. Save the bar plot as image and insert it here.
12. Save the bar plot as image and insert it here.
13. Report the p-value. What does it tell us about the significance of the difference in the number of positive tests between male and female populations?

If time permits, please continue to the next section.

**PART 3. Using animation to track COVID-19 infections over time**

**Introduction:**

After completing the first two parts, we will create an animation that can be used to visualize and track how COVID-19 infections change over time in each county. Such visualizations can potentially be used to help identify how the virus is spreading by showing when and where positive cases are most prominent.

To create this animation, we will generate a Graphical Interchange Format (GIF) file. A GIF is a series of images that will loop continuously. Specifically, the series of images that we will use will be the NJ heat maps created for each month. Thus, we will be able to see how the cases in each county are changing from month to month. However, in this exercise, the color scale that represents the number of positive cases per person in each county will be fixed for each month. The color scale must be fixed to the same high and low values for each month to compare how the positive cases are shifting from month to month. Using different values on the same color scale could lead to the misinterpretation of data and mislead the public.

The R package that will be used to create the GIF file from the ggplot heat maps is gifski. Read more about this package: <https://cran.r-project.org/web/packages/gifski/gifski.pdf>

If you cleared the environment from questions 1-13, import the saved environment by clicking on the folder icon with a green arrow and locating the saved NJ\_Covid.RData file. Once you have the script and the environment open in RStudio, you are ready to create the animation! Follow the instructions after the ### BREAK ### line in the script to create the animation and answer the following questions. For each Step, there is a task for you to complete to make the script work.

**Summary of Steps:**

STEP 3.1, import the GIF package

STEP 3.2, create unique heat map text naming labels

STEP 3.3, set working directory and create a new file to save images

STEP 3.4, use a for loop to automatically create and save each heat map from the data

STEP 3.5, organize and identify the saved heat map images

STEP 3.6, create the GIF

STEP 3.7, open the GIF

**Please answer the questions:**

14.

a) Based on the data in the animation, which county or counties appear to have a steady increase in their infection rate?

b) Observe the dynamics for Mercer county, Sussex county, and Camden County. Of the 3, which county was the best and the worse at containing the spread of COVID-19 from March through December? Explain your answer.

15.

a) Looking at the animation, do you think COVID-19 cases have spread from one location to other counties? Or do you think the infection spread from multiple locations at the same time in NJ? Explain what behavior in the animation led you to this hypothesis.

b) Philadelphia is West of Camden County while New York City is just to the East of Hudson county. Hypothesize how two highly densely populated cities may have impacted the spread of COIVD-19 in New Jersey.