Homework 3 and 4 – SEHBA WANI, 09.26.22

SECTION 1

Lyme disease is a tick-borne infection caused by the *Borrelia burgdorferi*. Incidence of Lyme disease have been increasing across the continental US, particularly in the Northeast. You have been given a dataset from the CDC of Lyme disease infection rates per 100,000 people by state.

1. *What was the overall infection rate in the US in 2008? What was it in 2015? What was it in 2019? (Hint: the US incidence is at the very bottom of each column.)*

   The overall infection rate for 2008 was 11.6 while for the year of 2019, it was 10.6. The overall infection rate for 2015 was 11.9 . The overall infection rate for 2019 was 10.6.

2. *List the 10 states with the highest rates of Lyme disease. (Hint: Use the arrange function. The grammar is the same as the filter function we used last week https://r4ds.had.co.nz/transform.html section 5.3.)*

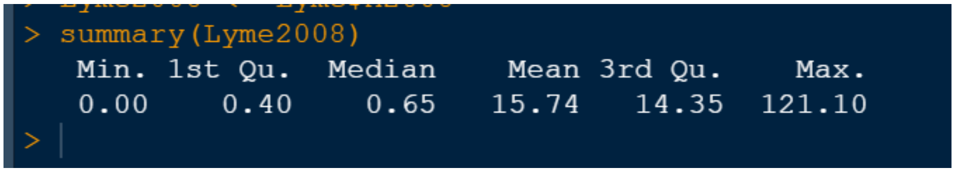
Vermont, Maine, New Hampshire, Rhode Island, Pennsylvania, Delaware, Connecticut, New Jersey, West Virgina, Wisconsion

3. *List the 10 states with the lowest rates of Lyme disease.*

The ten states with the lowest rates of Lyme disease can include, California, Arizona, Louisiana, Missouri , New Mexico , Texas, Colorado, Georgia , Mississippi, Hawaii, Oklahoma

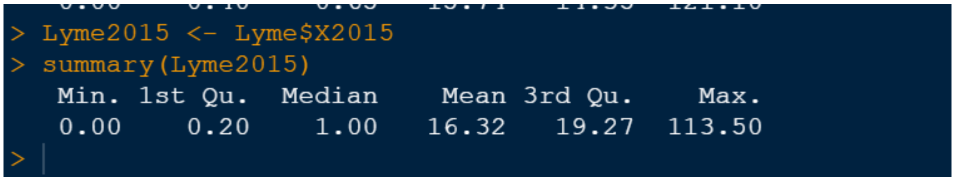
4. *Calculate summary statistics for 2008, 2015 and 2019. (Summary stats Helps kind of asks you what you are looking for the data, exploring data analysis)*

2008:



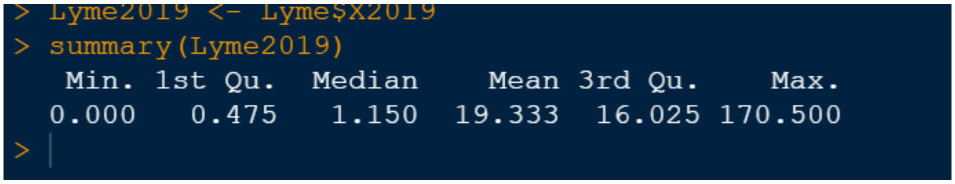
```
> Lyme2008 <- Lyme$X2008
> summary(Lyme2008)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    0.40    0.65   15.74   14.35  121.10 
> 
```

2015:



```
   0.00    0.40    0.65   15.74   14.35  121.10 
> Lyme2015 <- Lyme$X2015
> summary(Lyme2015)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
   0.00    0.20    1.00   16.32   19.27  113.50 
> 
```

2019:



```
> Lyme2019 <- Lyme$X2019
> summary(Lyme2019)
   Min. 1st Qu.  Median    Mean 3rd Qu.    Max. 
  0.000   0.475   1.150  19.333  16.025 170.500 
> 
```

5. Do you see any trends in the data based on the summary stats?

   Any trends that are noticeable in the data can include the central tendencies have increased. That says a lot about how the average has gone up, meaning that with time the cases of lyme disease has increased, instead of decreased. I also noticed that every year the minimum stayed the same, while the max increased many times.

<u>SECTION 2</u>

The gender inequality index (GII) provides insights into gender disparities in health, empowerment and the labor market. The GII is a composite measure, reflecting inequality in achievements between women and men in three dimensions: reproductive health, empowerment and the labor market.

- The health dimension is measured by the maternal mortality ratio and the adolescent fertility rate.
- The empowerment dimension is measured by the share of parliamentary seats held by each gender, and by secondary and higher education attainment levels.
- The labor dimension is measured by women's participation in the workforce.

The GII varies between 0 and with higher values representing more equal outcomes. It is designed to reveal the extent to which national human development achievements are eroded by gender inequality, and to provide empirical foundations for policy analysis and advocacy efforts.

1. *Import the data table and look at it. Do you see any issues you should deal with? (HINT: there are some missing values [https://r4ds.had.co.nz/transform.html section 5.2.3](https://r4ds.had.co.nz/transform.html))*

   When it comes to dealing with null values, there has to be a method as to how to deal with missing values, in this case that would be using the method of na.rm = TRUE.
2. *Create a histogram and a smoothed density estimate of the 2013 scores. Plot these on the same graph. (Doing the calculus – histogram then do the calculus draw the line)*

   Please look at my code regarding this question

3. *Now create a histogram and smoothed density estimate of the scores for 2010 and for 2008 as you did in question 2.*

    Please refer to code to see my histogram(s) and their smoothed densities.

4. *Calculate summary data for each year you have plotted. Have the means changed over time? How about the standard deviations? (Note: we will build on this when we learn t-tests.)*

    Summary stats are shown in the code.

The means and SDs have changed overtime. The means have increased, which may mean there are higher disparities of GI and the data is spread out more, showcasing and highlighting the impact of the average increasing overtime.

5. *Do the data look normally distributed? (we have not learned how to test for normality yet, so at this point it is an "eyeball" guess.*

The data does not look that normally distributed, there is a bit of left and right skew on each year. The smoothed density provides insight on the skewedness on the histogram.

<u>SECTION 3</u>

1. *Give three real world examples of binomial data.*
    1 - Heads or Tails
    2 - Binary (1's and 0's)
    3 – Yes and No
2. *Under what circumstances can a researcher authoritatively say that there is a causal link between two variables?*
    A researcher can say there is a causal link when there is a control group involved to eliminate confounding variables. As well as, if one variable change based on another variable.

3. *Develop a research and null hypothesis (Nothing is happening) for section 1. (has to be testable)*
    Null: The infection rate is not increasing, as time increases. Nothing is happening
    Research: The infection rate increases, as time increases.
4. *Develop a research and null hypothesis for section 2.*
    Null: As time increases, gender disparity does not decrease.
    Hypothesis: As time increases, gender disparity does decrease. i

Commented [SMW1]: I hope this isnt wrong and if it is ://///