Interquartile range differences among counties

Age distributions in Sao Paulo State

A box plot for Sao Paulo state (i.e. all of the counties in the spm.long.data.csv) by month was created in order to visualize the spread of age of infection throughout the course of the epidemic. This revealed very similar distributions of age for months 5-10 with an average age of infection of approximately 20 years old. The box plot also revealed that the majority of outliers occurred in months 7, 8, and 9.

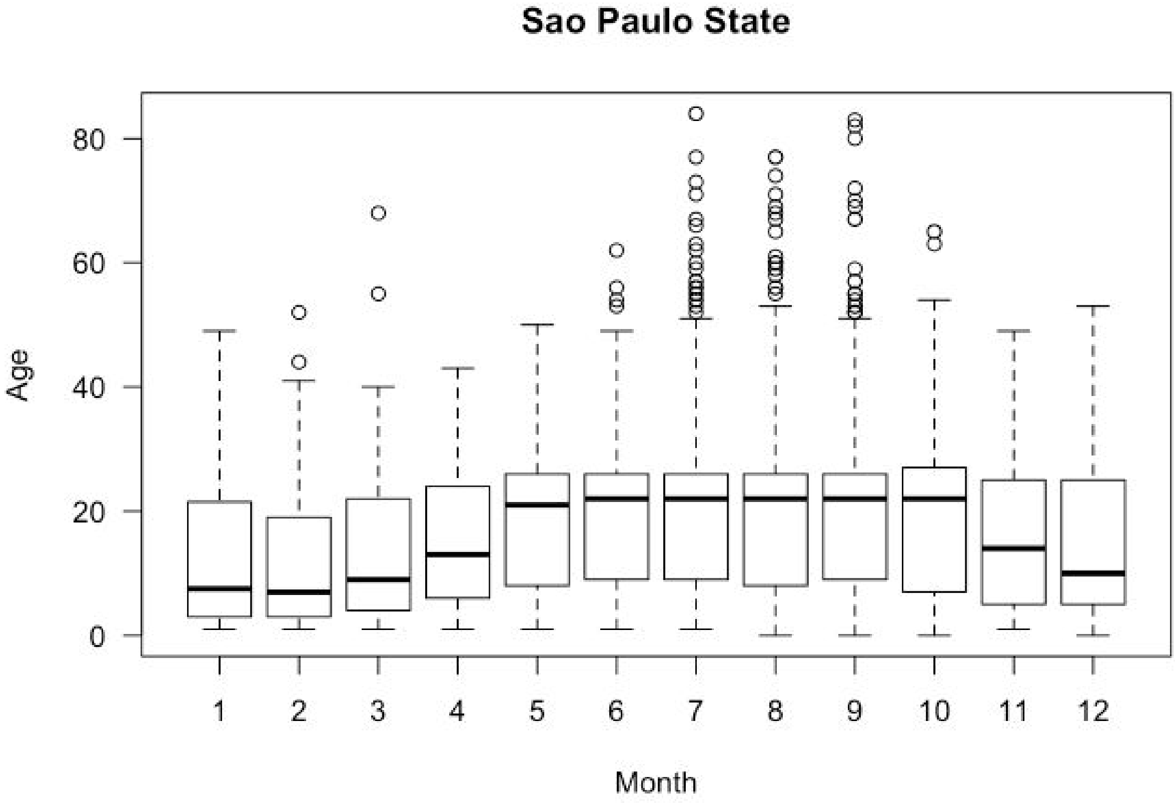


Figure 1. Age distribution of cases throughout the epidemic. As seen by the boxplots in months 5-9, the average age of infection was approximately 20 years old. Additionally, as indicated by the small difference between the mean and the 75 percentile, there was a large density of cases from ~20 to ~25 years of age during months 5-9.

Descriptive statistics of quartile ranges

It was also found that the average lower quartile was 12.67 years old with a standard deviation of 9.6. The average upper quartile was 18.85 years old with a standard deviation of 10.43.

Relation between interquartile range and month and interquartile range and county

An ANOVA was ran between county and interquartile range in counties with more than 20 cases. The null hypothesis that county and interquartile range are unrelated can be rejected with a p = 0.00513. Another ANOVA was ran between month and interquartile range in counties with more than 20 cases and the null hypothesis that month and IQR are unrelated can be rejected with $p = < 2 \times 10^{-16}$.

Upper quartile and RMSP

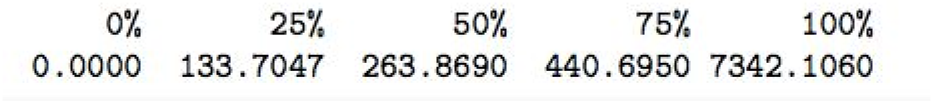
An ANOVA was run between the upper quartile range for each county and RMSP. RMSP is a variable which indicates whether the county in question is (RMSP = 1) or is not (RMSP = 0) in the greater regional Sao Paulo area. The residuals of this model violated the assumption of normality (Shapiro-Wilk normality test yielded a p-value $< 2.2 \times 10^{-16}$). No effective transformations were found, so the nonparametric Kruskal-Wallis test was ran and yielded a p-value of 0.0008138. This allows for acceptance of the alternative hypothesis that RMSP affects the upper quartile value.

Descriptive statistics were also found for the upper quartile values and it was found that the mean of upper quartiles for those within the regional Sao Paulo area was 64.165493 with a standard deviation of 28.9415899. For counties outside of the regional Sao Paulo area, the mean upper quartile value was 73.468106 with a standard deviation of 40.3572421.

Lower quartile and RMSP

The same analysis was run to evaluate RMSP's effect on the lower quartile. The ANOVA yielded insignificant results (p = 0.598), but also violated the assumption of normality (Shapiro-Wilk normality test yielded a p-value of $< 2.2 \times 10^{-16}$). The nonparametric Kruskal-Wallis test was ran and yielded an insignificant p-value of 0.2159. Upon preliminary inspection, the null hypothesis that RMSP does not affect the lower quartile must be accepted.

In order to more closely examine this, the distance between each county and Sao Paulo county was calculated using the package "ggmap." The quartile values of distance from Sao Paulo county was found to be the following:



```
       0%       25%       50%       75%      100%
   0.0000  133.7047  263.8690  440.6950 7342.1060
```

These values were utilized in order to create four distance classes to serve as factors in an ANOVA. A fifth class was created in order to encompass those counties which the distance from Sao Paulo county could not be found. The ANOVA between lower quartile and distance class yielded non-normal, insignificant results (SW p-value $< 2.2 \times 10^{-16}$ and ANOVA p-value = 0.111). A nonparametric Kruskal-Wallis test was ran and yielded a p-value of 0.08664.

Because the KW test yielded a relatively low insignificant p-value, further inspection of the analysis was necessary. The entries in the lower quartile values which were a part of distance class "E" (i.e. the distance between the county in interest and Sao Paulo county could not be found), were removed from the dataset. Another ANOVA was ran and yielded non-normal, insignificant results (SW p-value < 2.2×10^{-16} and ANOVA p-value = 0.143). A nonparametric, KW test also yielded insignificant results with p = 0.1025. With these results in mind, it appears that distance from Sao Paulo county does not affect the lower quartile age of infection.

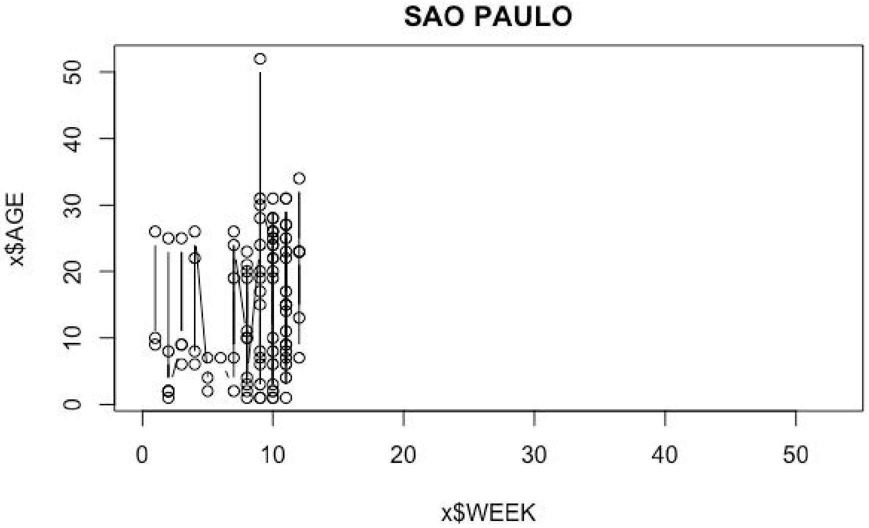
Identification of spark cases

First one hundred cases in a county

In order to identify the pattern of initial spread to new counties, the first one hundred cases were examined. Cases which brought initial disease were termed “spark cases.” It was a goal to identify any patterns among these “sparkers.” In line with the immigration hypothesis, one should expect these first one hundred cases to display a pattern of adults leading the infection pathway in both urban and rural counties.

The matrix “index_week_mat” was created and consists of the index, age, county, and week of the first one hundred cases in each county. Where counties had less than one hundred cases throughout the entire course of the outbreak, all cases were concluded. This was then sorted by week in ascending order. Scatter plots for each county which displayed the age (y-axis) by week (x-axis).

These scatter plots are not the most intuitive/informative graphics for several reasons. In counties in which the outbreak took off relatively fast, the graphs are very condensed and do not provide much information. This is displayed in the Sao Paulo graph below.



As can also be seen in the Sao Paulo graph, when multiple cases occur in the same week, they appear on the graph as a vertical line. Additionally, a consequence of utilizing a line graph is the implication of causation between cases in the previous week and cases in the current week (the non-vertical diagonal line from week 6 to week 7 seems to imply that the cases in week 6 caused the cases in week 7).

The analysis of the first one hundred cases was going to be supplemented by analysis by district, however it was discovered that district is only listed for the cases in Sao Paulo county.

The only true take-away from analysis of the first one hundred cases was a visual inspection of when, and how long it took for the first one hundred cases to occur. Ultimately, this analysis did not prove to be of much use.

Stage of epidemic

The stage of the epidemic relative to the stages of epidemic in other counties was quantified in order to approximate the likelihood of the epidemic being brought from one county to another. This was done by estimating two factors in disease spread on a county level: the county's infectedness (how many people were infected at one time) and it's sociability.

A county's infectedness was calculated by assessing how many cases were active during any given week. This was done using the function active_cases(*i, j*), with *i* being the index of the county in county_vec and *j* being the week of the epidemic. This was then scaled by the county's population in order to assess the proportion of the population that infected individuals accounted for. The scaled version of active_cases(*i, j*) is prop.active_cases(*i, j*).

In order to make a county's infectedness reflect the immigration hypothesis, I have also played with making prop.active_cases(*i, j*) sensitive to the average age of infection in that county at that week. This would hopefully reflect that those with an older age of infection should be more likely to spread the disease (because they contain unvaccinated adults). I propose two ways to do this in my code, and have listed them here (sf = scaling factor):

1. sf = mean age / max age
2. sf = mean age

I am somewhat hesitant to continue on this course because I am not sure how reflective it is of critical community size.

A county's sociability was quantified in order to determine how likely two counties were to interact. In order to do this, the function distance(*k, i*) was created to find the distance from county_vec[*i*] to county_vec[*k*]. The distance between two counties was then used in order to create social(*k, i*). This function quantifies the interaction between the two counties as the following:

1 - (distance(*k, i*) / sqrt(area of Sao Paulo state))

These two characteristics (infectedness and sociability) of the counties were multiplied together using the function spread(*i, j, k)* which hypothetically quantifies the chance that the infection spread from county *i* to county *k* in week *j*.

This function was then used to test whether unvaccinated rural adults were acting as spreaders of disease to urban areas. This is depicted when county *i* is rural and county *k* is urban. I plan to assess this by completing a one-tailed t-test with the following hypotheses:

H_o: $\mu_{\text{rural to urban}} = \mu_{\text{urban to rural}}$

H_o: $\mu_{\text{rural to urban}} > \mu_{\text{urban to rural}}$