

Lower Quartile Differences between Urban and Rural

Maggie Walters

June 18, 2017

Document synopsis

The differences in lower quartile between urban and rural communities is being explored because it was found that while RMSP has an effect on upper quartile, it does not on the lower. This observation will be explored in the following ways:

- Change RMSP to have more factors by creating distances from Sao Paulo using the Google maps package.
 - Possibly do five distance groups: 0-20 km, 20-40 km, 40-60 km, 60-80 km, and 80+ km.
- If the null hypothesis is rejected for this, do a MCT (can these be done for non-parametric tests?) in order to see which distance groups shows a significant effect on lower quartile.

Finding Distance between each County and Sao Paulo County

The package “ggmap” was used in order to find the distance between each county and Sao Paulo county in kilometers. This will be used to make distance classes that are more specific than RMSP.

```
## Installing package into '/Users/maggiewalters/Library/R/3.3/library'
## (as 'lib' is unspecified)

## Warning: unable to access index for repository YOUR FAVORITE MIRROR/src/contrib:
##  scheme not supported in URL 'YOUR FAVORITE MIRROR/src/contrib/PACKAGES'

## Warning: package 'ggmap' is not available (for R version 3.3.3)

## Warning: unable to access index for repository YOUR FAVORITE MIRROR/bin/macosx/mavericks/contrib/3.3
##  scheme not supported in URL 'YOUR FAVORITE MIRROR/bin/macosx/mavericks/contrib/3.3/PACKAGES'

## Loading required package: ggplot2

## by using this function you are agreeing to the terms at :
## http://code.google.com/apis/maps/documentation/distancematrix/
## Information from URL : http://maps.googleapis.com/maps/api/distancematrix/json?origins=SAO+PAULO&des
## by using this function you are agreeing to the terms at :
## http://code.google.com/apis/maps/documentation/distancematrix/
## Information from URL : http://maps.googleapis.com/maps/api/distancematrix/json?origins=SAO+PAULO&des
## by using this function you are agreeing to the terms at :
## http://code.google.com/apis/maps/documentation/distancematrix/
## Information from URL : http://maps.googleapis.com/maps/api/distancematrix/json?origins=SAO+PAULO&des
## by using this function you are agreeing to the terms at :
## http://code.google.com/apis/maps/documentation/distancematrix/
## Information from URL : http://maps.googleapis.com/maps/api/distancematrix/json?origins=SAO+PAULO&des
## by using this function you are agreeing to the terms at :
```

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

[illegible]

Creating Distance Classes

Distance from Sao Paulo county was classified into 5 age group classes using quantile classifications, as followed:

- A: $0 < x < 133.7047$
- B: $133.7047 < x < 263.869$
- C: $263.869 < x < 440.695$
- D: $440.695 < x < 7342.106$
- E: NA

These were then converted into factors in order to perform an analysis of variance with the following null and alternative hypotheses:

H_0 : Distance class does not affect the lower quantile value. H_A : Distance class does affect the lower quantile value.

```
DISTANCE_IN_KM <- as.numeric(DISTANCE_IN_KM)
quantile(DISTANCE_IN_KM, na.rm = TRUE)

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

distance_class <- rep(NA, length(DISTANCE_IN_KM))

which.A <- which(0 <= DISTANCE_IN_KM & DISTANCE_IN_KM <= 133.7047)
which.B <- which(133.7047 < DISTANCE_IN_KM & DISTANCE_IN_KM <= 263.8690)
which.C <- which(263.8690 < DISTANCE_IN_KM & DISTANCE_IN_KM <= 440.6950)
which.D <- which(440.6950 < DISTANCE_IN_KM & DISTANCE_IN_KM <= 7342.1060)
which.E <- which(is.na(DISTANCE_IN_KM))

distance_class[which.A] <- "A"
distance_class[which.B] <- "B"
distance_class[which.C] <- "C"
distance_class[which.D] <- "D"
distance_class[which.E] <- "E"
```

Setting up table for ANOVA

Code taken from “descriptive_statistics_by_urba:rural.Rmd”

```
#create blank lower quartile matrix
lowerq_matrix <- matrix(rep(NA, length(county_vec) * 12), nrow = 12, ncol = length(county_vec))
colnames(lowerq_matrix) <- county_vec
months <- c("January", "February", "March",
            "April", "May", "June",
            "July", "August", "September",
            "October", "November", "December")
row.names(lowerq_matrix) <- months

#find lower quartile for each county in each month
for(i in 1:12){
  x <- subset(data, data$MONTH == i)
  for(j in 1:387){
    which.county_j <- which(x$COUNTY == county_vec[j])
    county_j_ages <- x$AGE[which.county_j]
    lowerq_matrix[i,j] <- quantile(county_j_ages)[2]
  }
}

lower_matrix <- matrix(rep(NA, 4 * 12*length(county_vec)), ncol = 4)
colnames(lower_matrix) <- c("LOWER", "COUNTY", "MONTH", "DISTANCE")

#fill in lower
for(i in 1:12 * length(county_vec)){
  lower_matrix[,1] <- lowerq_matrix[,]
```

```

}
#fill in months
lower_matrix[,3] <- rep(seq(1:12), length(county_vec))
#fill in counties
lower_matrix[seq(1,12),2] <- rep(county_vec[1], 12)
for(i in 2:387){
  x <- 12 * (i-1)
  lower_matrix[x + seq(1,12),2] <- rep(county_vec[i], 12)
}
#fill in distance class
lower_matrix[seq(1,12),4] <- rep(distance_class[1],12)
for(i in 1:length(distance_class)){
  x <- 12 * (i-1)
  lower_matrix[x + seq(1,12),4] <- rep(distance_class[i], 12)
}

lower_matrix <- as.data.frame(lower_matrix)
lower_matrix <- subset(lower_matrix, !is.na(lower_matrix$LOWER))

```

Run ANOVA or NP test

```

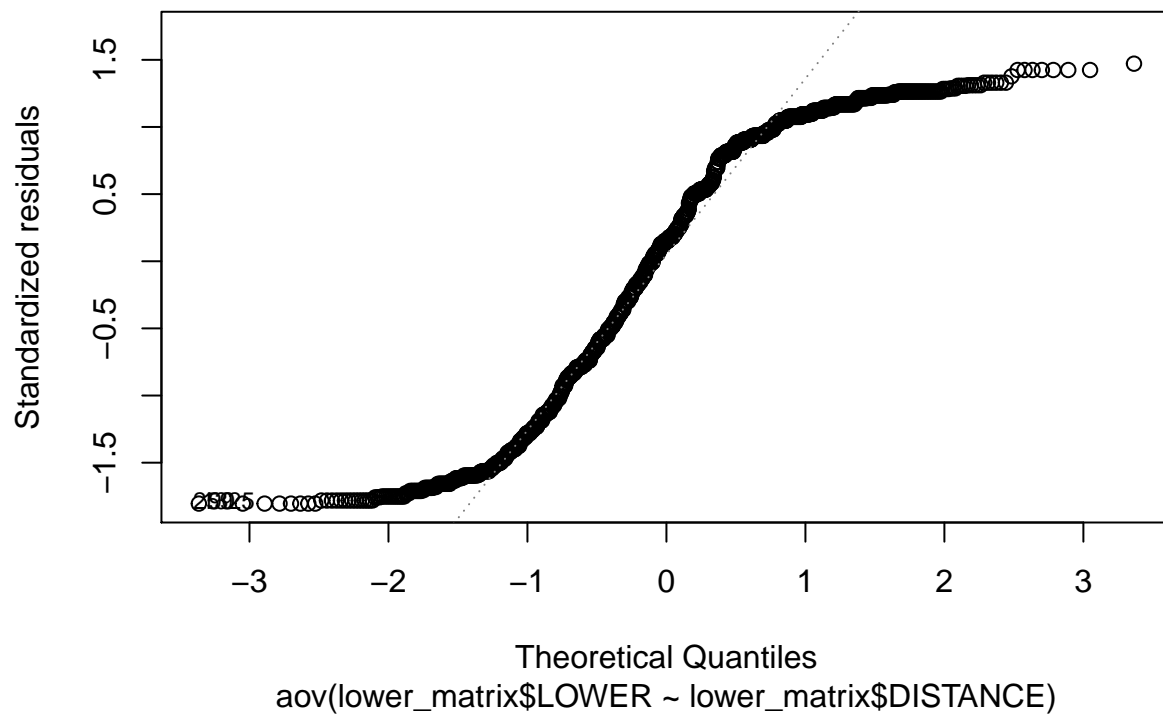
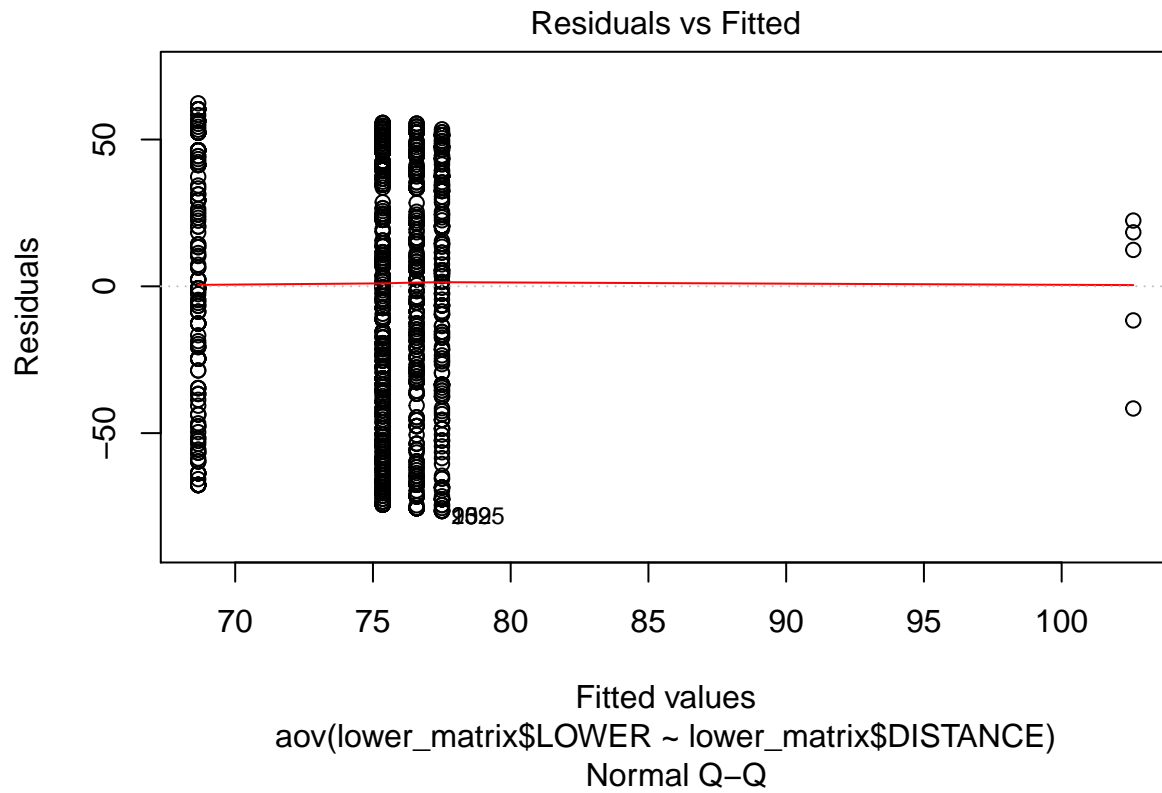
lower_matrix$LOWER <- as.numeric(lower_matrix$LOWER)

#ANOVA, p = 0.111
lower_distance_class_mod <- aov(lower_matrix$LOWER ~ lower_matrix$DISTANCE, data = lower_matrix)
#assess for normality, NOT NORMAL
shapiro.test((resid(lower_distance_class_mod)))

##
## Shapiro-Wilk normality test
##
## data: (resid(lower_distance_class_mod))
## W = 0.91351, p-value < 2.2e-16

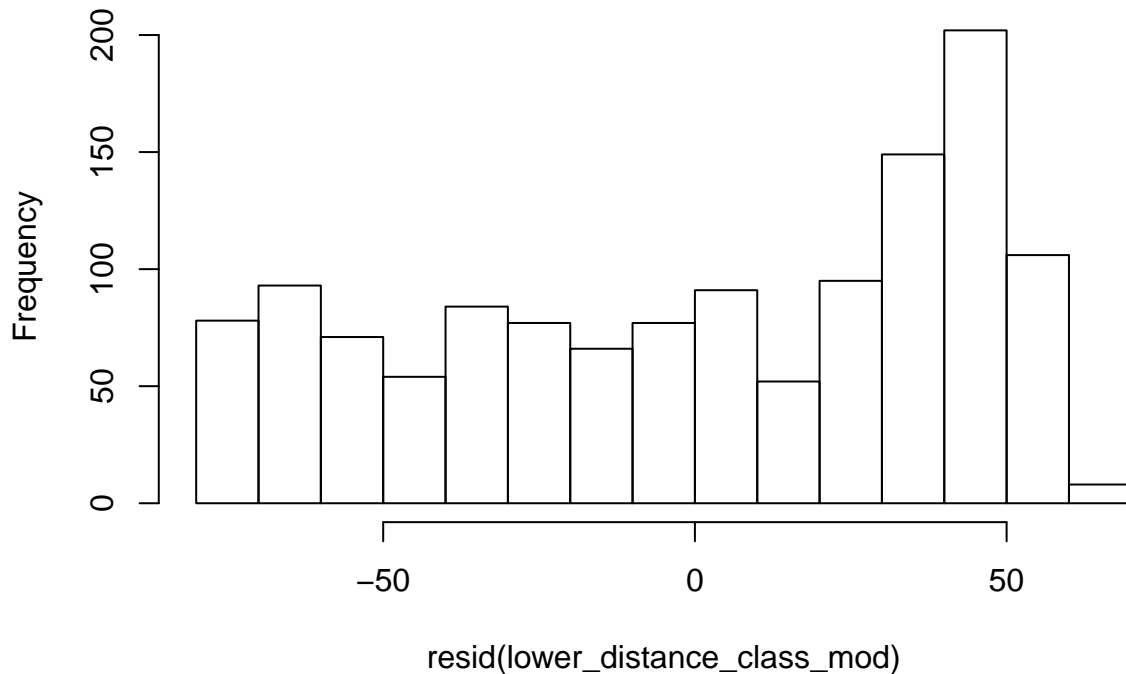
plot(lower_distance_class_mod, which = c(1,2))

```



```
hist(resid(lower_distance_class_mod))
```

Histogram of resid(lower_distance_class_mod)



```
#NON-PARAMETRIC TEST, p = 0.08664
np_lower_distance_mod <- kruskal.test(lower_matrix$LOWER ~ lower_matrix$DISTANCE, data = lower_matrix)
```

Preliminary analysis:

The ANOVA for assessing lower quantile with distance class yielded insignificant ($p = 0.111$) and non-normal results. This led to running a non-parametric test (Kruskal-Wallis), which also yielded insignificant results ($p = 0.08664$). This leads to failure to reject the null hypothesis. However, There were three counties which had a NA value because they were not found via a preliminary Google search. Moving forward, I will remove these three counties in order to see if this leads to significant results.

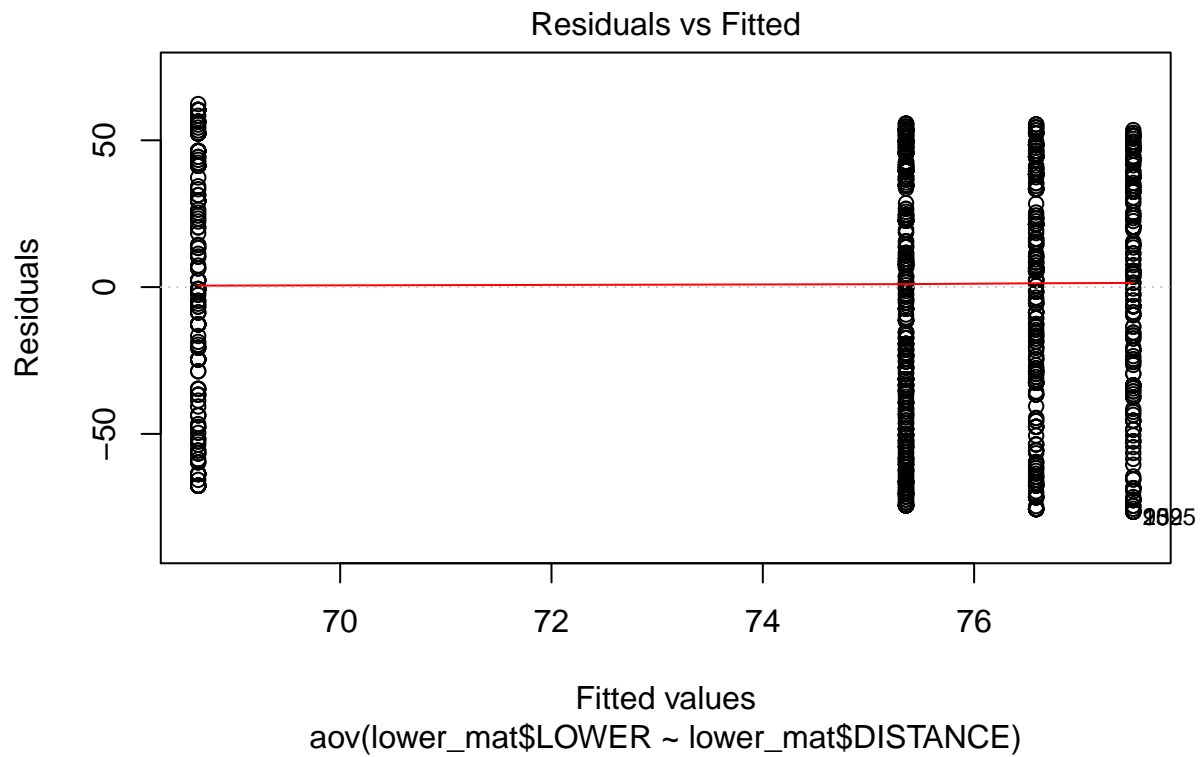
ANOVA/ NP tests without NA listings

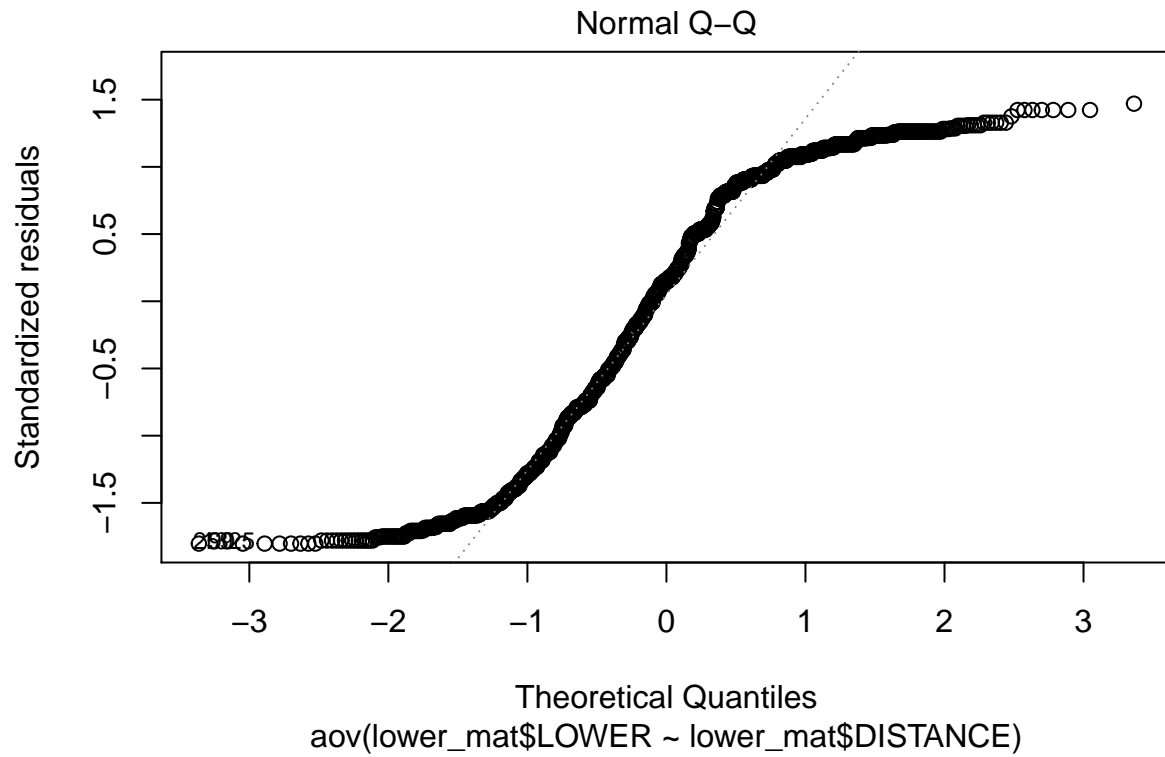
```
which.E.mat <- which(lower_matrix$DISTANCE == "E")

lower_mat <- lower_matrix[~1115,]
lower_mat <- lower_mat[~889,]
lower_mat <- lower_mat[~888,]
lower_mat <- lower_mat[~887,]
lower_mat <- lower_mat[~753,]

#ANOVA without the NA's
lower_mat$LOWER <- as.numeric(lower_mat$LOWER)
#p = 0.143
lower_mat_aov <- aov(lower_mat$LOWER ~ lower_mat$DISTANCE, data = lower_mat)
#assess for normality, NOT NORMAL
shapiro.test((resid(lower_mat_aov)))
```

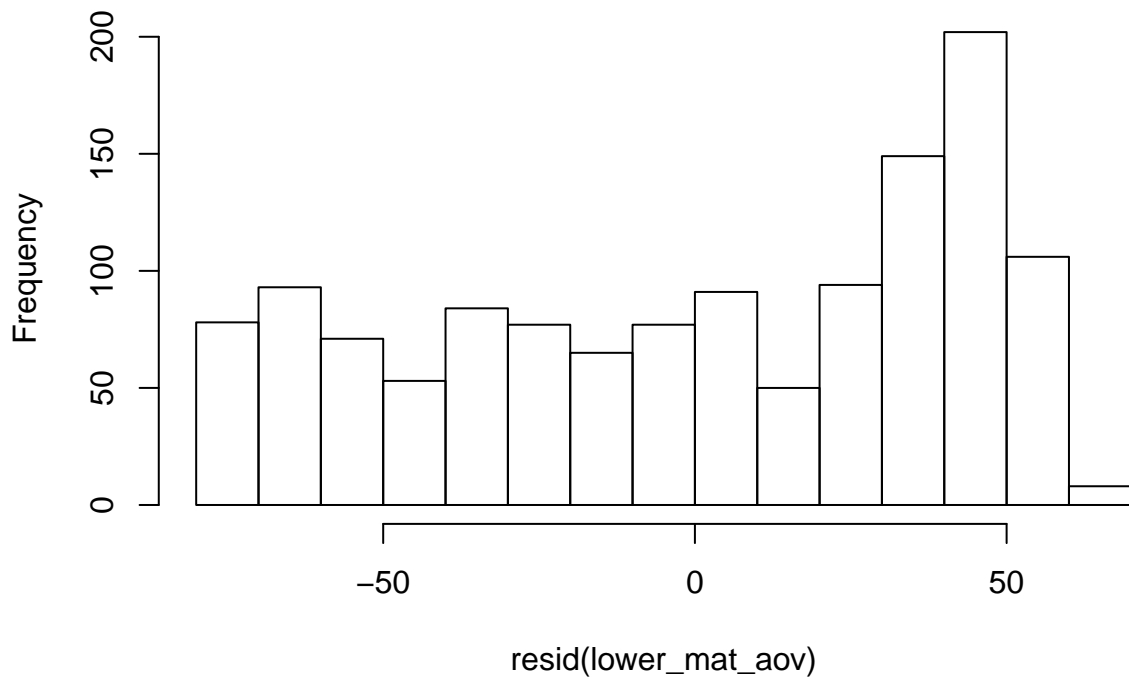
```
##
## Shapiro-Wilk normality test
##
## data:  (resid(lower_mat_aov))
## W = 0.91299, p-value < 2.2e-16
plot(lower_mat_aov, which = c(1,2))
```





```
hist(resid(lower_mat_aov))
```

Histogram of resid(lower_mat_aov)



```
#NON-PARAMETRIC TEST, p = 0.1025
np_lower_mat <- kruskal.test(lower_mat$LOWER ~ lower_mat$DISTANCE, data = lower_mat)
```

Preliminary analysis

Obtained a higher p-value for both of the analysis of variance tests (with $p = 0.143$ for the ANOVA and $p = 0.1025$ for the non-parametric test). Again, the ANOVA failed to achieve normality. I feel that at this point then, that the null hypothesis must be accepted. Essentially, **distance from Sao Paulo county does not affect the lower quartile of age of infection.**