**Data Analytics Assignment 3 Write-Up**

COVID-19 us-counties datasets used: us-counties-2020, us-counties-2021

Note: To have the plots be meaningful, the data was subset to remove most outliers past the 3rd quartile. Thus, cases data for us-counties-2020 was subset to only include values below 2000, cases data for us-counties-2021 was subset to only include values below 8500, deaths data for us-counties-2020 was subset to only include values below 75, and deaths data for us-counties-2021 was subset to only include values below 200.

A diagram of a box plot

Description automatically generated

A graph showing the number of deaths in a box plot

Description automatically generated

Boxplot summary statistics for COVID-19 Cases for All U.S. Counties:

A screenshot of a computer screen

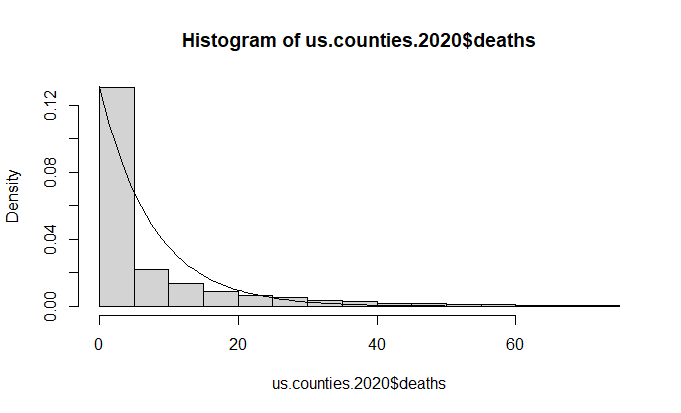
Description automatically generated

Boxplot summary statistics for COVID-19 Deaths for All U.S. Counties:

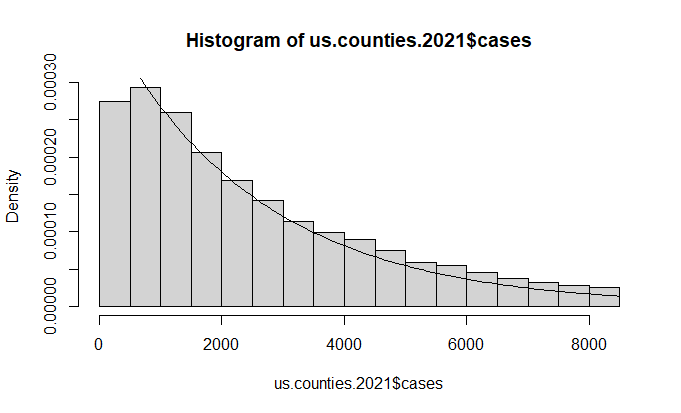
A screenshot of a computer screen

Description automatically generated

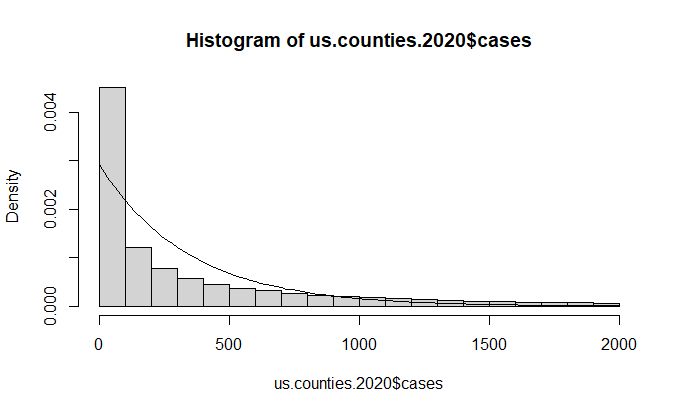
1. Looking at the summary statistics for the boxplot generated for the 2020 and 2021 COVID-19 cases, there was a significant increase in cases at all levels from 2020 to 2021, with the third quartile for 2020 cases being less than the first quartile for 2021 cases. For deaths, while the numbers were much smaller than cases, there was also a significant increase from 2020 to 2021, with again the third quartile for 2020 deaths being less than the first quartile for 2021 deaths. Looking at the box plots themselves, there were also lots of outliers for each dataset, as certain counties and states had spikes in COVID-19 cases and deaths.



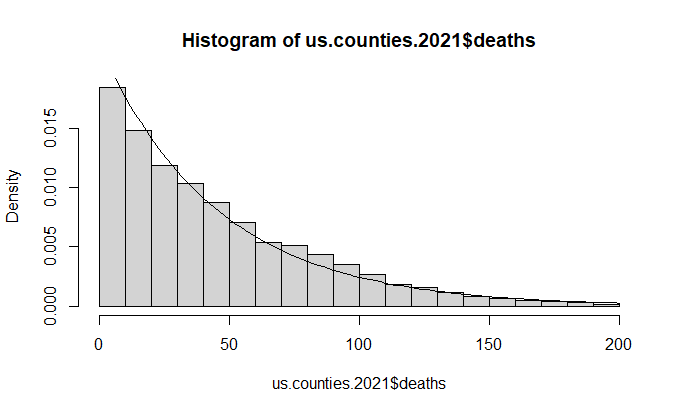
Curve used: curve(dexp(x,1/mean(us.counties.2020$cases)), add=TRUE)



Curve used: curve(dexp(x,1/mean(us.counties.2021$cases)), add=TRUE)



Curve used: curve(dexp(x,1/mean(us.counties.2020$deaths)), add=TRUE)



Curve used: curve(dexp(x,1/mean(us.counties.2021$deaths)), add=TRUE)

1. The histograms for the Cases and Deaths variables in each of the datasets mainly resembled an exponential distribution, as most of the counties had low numbers of Casea and Deaths, with a steep drop-off in the number of counties with higher Cases and Deaths. I did also explore other distributions, like normal, gamma and t-distributions, however they didn’t seem to match the histogram as much, with gamma (when setting shape to a low number and scale to a high number) being the second closest fit. I decided to go with the exponential distribution as the distribution fit for each of the histograms.

A graph of a graph

Description automatically generated with medium confidence

A graph of a graph of a graph

Description automatically generated with medium confidence

A graph of a line

Description automatically generated with medium confidence

Q-Q plot for COVID-19 Cases for All U.S. Counties in 2020, against Exponential Distribution

A graph of a curve

Description automatically generated

Q-Q plot for COVID-19 Cases for All U.S. Counties in 2021, against Exponential Distribution

A graph of a curve

Description automatically generated

Q-Q plot for COVID-19 Deaths for All U.S. Counties in 2020, against Exponential Distribution

A line graph with numbers and points

Description automatically generated

Q-Q plot for COVID-19 Deaths for All U.S. Counties in 2021, against Exponential Distribution

1. The ECDF plots each comparing Cases and Deaths between 2020 and 2021 show that, in 2020, most of the cases and deaths were low in number, while in 2021, the number of cases and deaths were more spread out. The Q-Q plots also show that the exponential distribution selected in 1b were a good fit, with the Q-Q points resembling a straight line, but still with some outliers.

A graph of death in california

Description automatically generated

A graph of a person with a graph

Description automatically generated with medium confidence

A graph of cases with numbers and text

Description automatically generated

A graph of death

Description automatically generated

A graph of a number of cases

Description automatically generated

A graph of a graph

Description automatically generated

A graph of a number of points

Description automatically generated

A graph of a line

Description automatically generated with medium confidence

A line graph with numbers and dots

Description automatically generated

A graph of a line

Description automatically generated with medium confidence

1. I decided to filter the data by a single state (which has numerous counties), California, to see how the histograms, ECDF plots, and Q-Q plots would be generated for the reported number of cases and deaths for this state in years 2020 and 2021. All trends remained consistent, so that was a big similarity. For differences, the histograms showed that a higher percentage of counties had low numbers of cases and deaths, but with outliers that were very far away. The ECDF plots featured slightly less smooth curves, but overall still showed that 2021 had a higher variance in cases and deaths than 2020. The Q-Q plots showed the exponential distribution fitting the data, like in question 1, but there were much bigger outliers this time.

A graph of two people

Description automatically generated with medium confidence

Bath vs. Price for NY Houses

Note: Does not represent all data captured, some outliers are omitted.

A graph of black dots

Description automatically generated with medium confidence

Residuals for Linear Model for Price of NY Houses

Note: Does not represent all data captured, some outliers are omitted.

Linear Model Summary:

A screenshot of a computer

Description automatically generated

1. The variable most significantly influencing house price is BATH, as it has the highest coefficient in the linear regression model. However, when plotting it against the PRICE, it shows that price fluctuates quite a lot for any number of baths, with some prices being very high. When plotting the best fit line for the linear model for price, it in fact shows a line with negative slope, which shows that even though BATH has the highest coefficient, the other variables have large enough coefficients to also affect the price, even affecting the price in the opposite direction compared to baths. The residuals show that the predicted values of the linear model differ greatly from the observations, mostly up to $1,000,000.

A graph of a line

Description automatically generated with medium confidence

Bath vs. Price for NY Houses (below $1,000,000 in Price)

Note: Does not represent all data captured, some outliers are omitted.

A graph of black dots

Description automatically generated with medium confidence

Residuals for Linear Model for Price of NY Houses (below $1,000,000 in Price)

Note: Does not represent all data captured, some outliers are omitted.

Linear Model Summary:

A screenshot of a computer

Description automatically generated

1. I repeated the linear model with a subset of the data for house prices below $1,000,000. I got similar results as in 3a, except the slope for the best fit line from the linear model for price is now positive, implying that the number of baths seems to have a greater impact on houses priced below $1,000,000. However, the plotted baths against price still show a massive fluctuation in prices. Additionally, the residuals show differences mostly up to $400,000.