Comparing COVID-19 Deaths Reported by Age Group

By Christopher Clegg

ece.clegg@gmail.com

Why? : Learning to Extract Meaningful Insights From Data

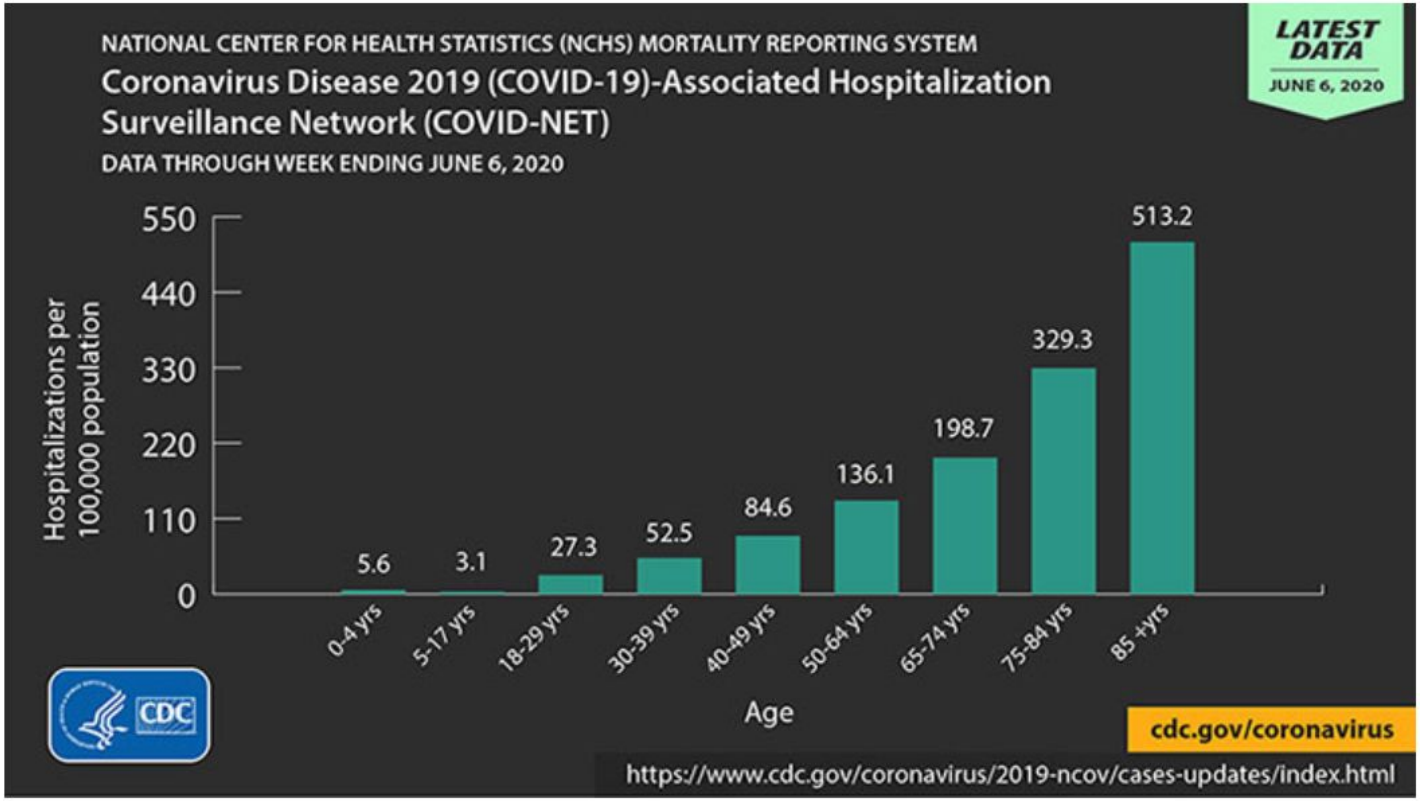
- **The motivation:**
  - Practice my new **data science** skills
  - Interesting, important, and worth knowing
- **The data:** deaths attributed to the **COVID-19** virus, aka: the **Coronavirus**, aka: **SARS-CoV-2** (severe acute respiratory syndrome coronavirus 2)

According to the CDC website:

> “As you get older, your risk for severe illness from COVID-19 increases. For example, people in their 50s are at higher risk for severe illness than people in their 40s. Similarly, people in their 60s or 70s are, in general, at higher risk for severe illness than people in their 50s. The greatest risk for severe illness from COVID-19 is among those aged 85 or older.”

Last Updated Sept. 11, 2020
Content source: National Center for Immunization and Respiratory Diseases (NCIRD), Division of Viral Diseases
Found at:
https://www.cdc.gov/coronavirus/2019-ncov/need-extra-precautions/older-adults.html#:~:text=The%20greatest%20risk%20for%20severe,as%20having%20underlying%20medical%20conditions.

Will the reported deaths impact these age groups the same way the reported hospitalization have?



Forming a black and white question:

When I consider all the records for deaths attributed to COVID-19 that had at least one death, can a statistically significant difference be observed between age groups, and how does this compare with the reported deaths attributed to the ‘Flu’ or Pneumonia?

Doctors and medical professionals in the media claim that the elderly are at a greater risk of having a fatal encounter with COVID-19. Does this data set confirm or contradict that claim?

- **The null hypothesis**

  (The difference of (means or medians) is ~0)

- **The alternative hypothesis**

  (or prediction that the (means or medians) will have a statistically significant difference)

- What do you think will happen?
- Will the null be rejected and the alternative be accepted, or will the null be kept?

Research: A brief “play-by-play”

- Search for data
  - Found some open source data at: https://data.cdc.gov/NCHS/Provisional-COVID-19-Death-Counts-by-Sex-Age-and-S/9bhg-hcku
  - Alway read accompanying documentation
- Exploring and Cleaning
  - Data rarely comes in ideal shape and never comes conveniently pre-arranged just for you.
  - Determine arrangement and eliminate overlap and excess.

Research: A brief “play-by-play” cont’d...

- Getting a feel for the data
  - Preliminary visualization:
    - Distribution
- What Tools Should I Use?
  - Determine testing methods
  - Which test best fits the distribution of my data sets?

Research: A brief “play-by-play” cont’d...

- Apply test methods
- Do post hoc analysis
- Visualize the results and put them into an explanation that is relatable and form a final conclusion

My methods

These are the method I chose, and why I chose them.

Not your normal distribution...

- Histogram - skewness, kurtosis
- How many are being compared?
- Not normal in the same way?
- Quartile/quartile plot, box plot
- Kruskal-Wallis Test
- Confidence Interval, Common Language Effect size

Distribution of COVID-19 Deaths.

Each color represents a different Age group.

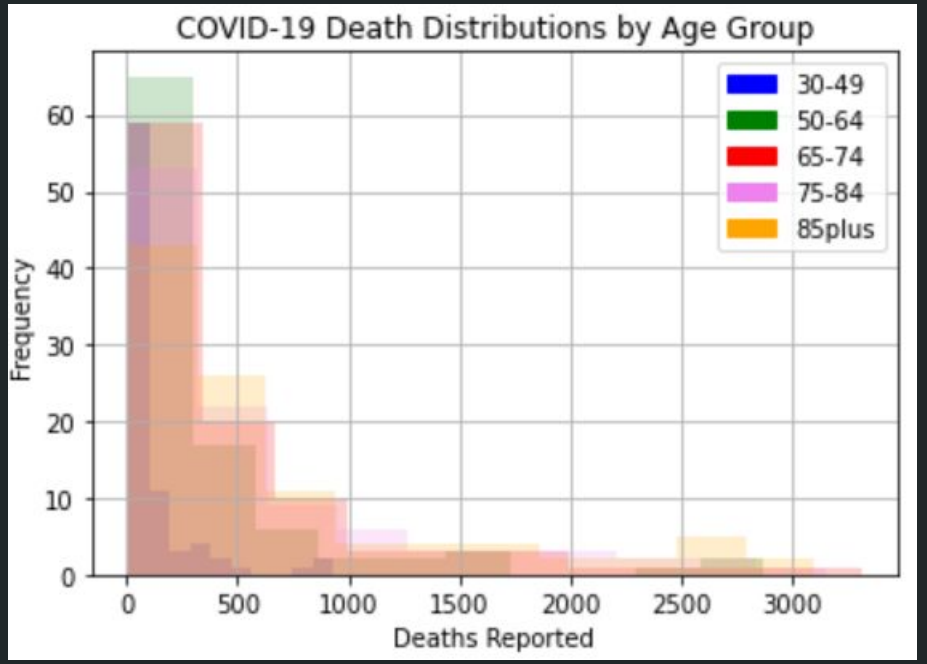


Figure 1. Each of the age groups are similarly distributed

Distribution of Pneumonia Deaths.

Each color represents a different age group

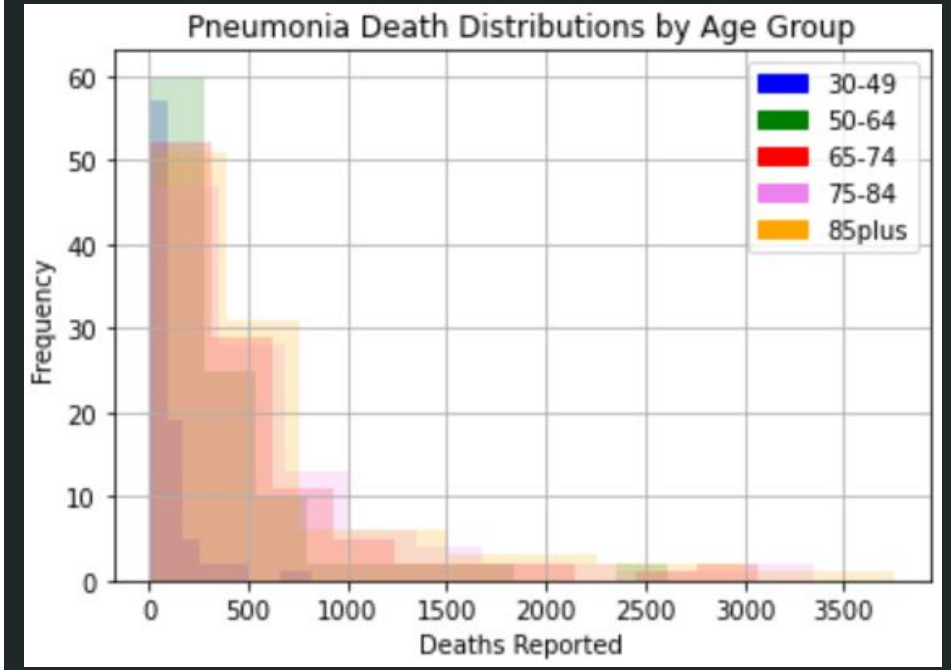


Figure 2. Each of the age groups are similarly distributed

Distribution Deaths in the 0-49 Age Group For COVID-19 and Pneumonia.

Each color represents a different age group

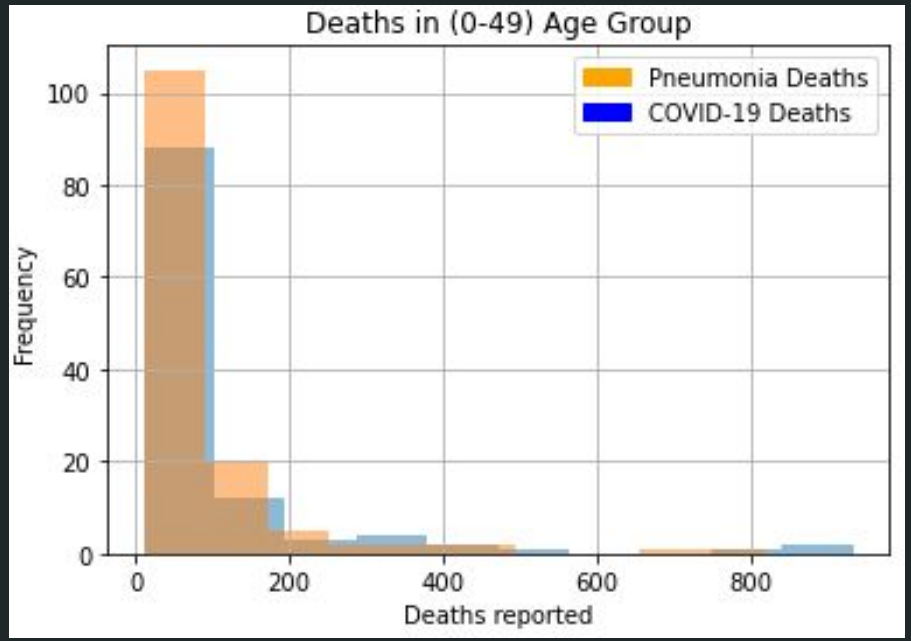


Figure 3. Each are similarly distributed

Distribution Deaths in the 50+ Age Group For COVID-19 and Pneumonia.

Each color represents a different age group

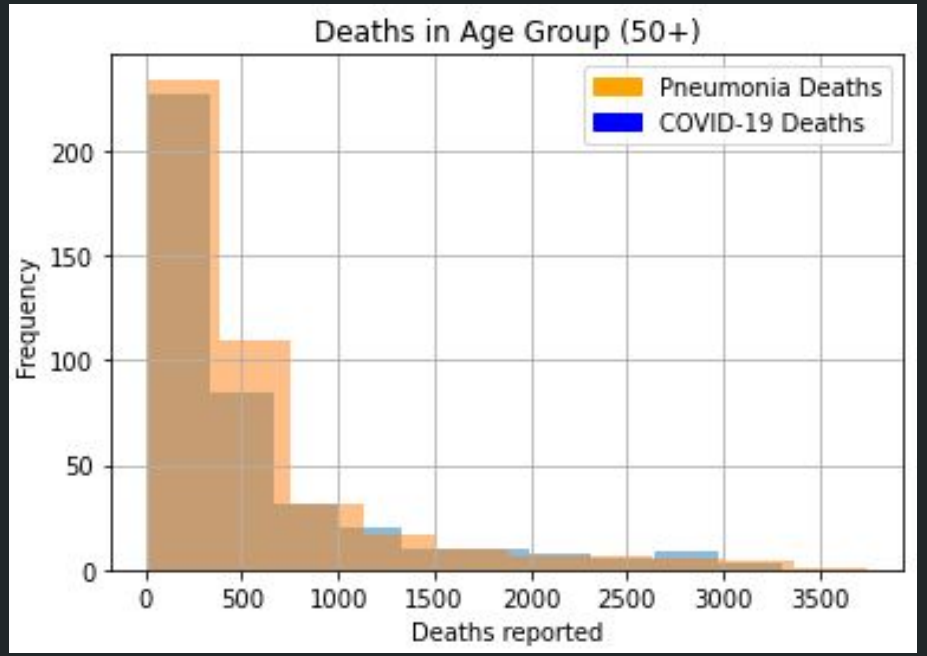


Figure 4. Each are similarly distributed

Quartile-Quartile Plots.

Red line represents normality
Both data sets are ‘non-normal’ in a similar way.

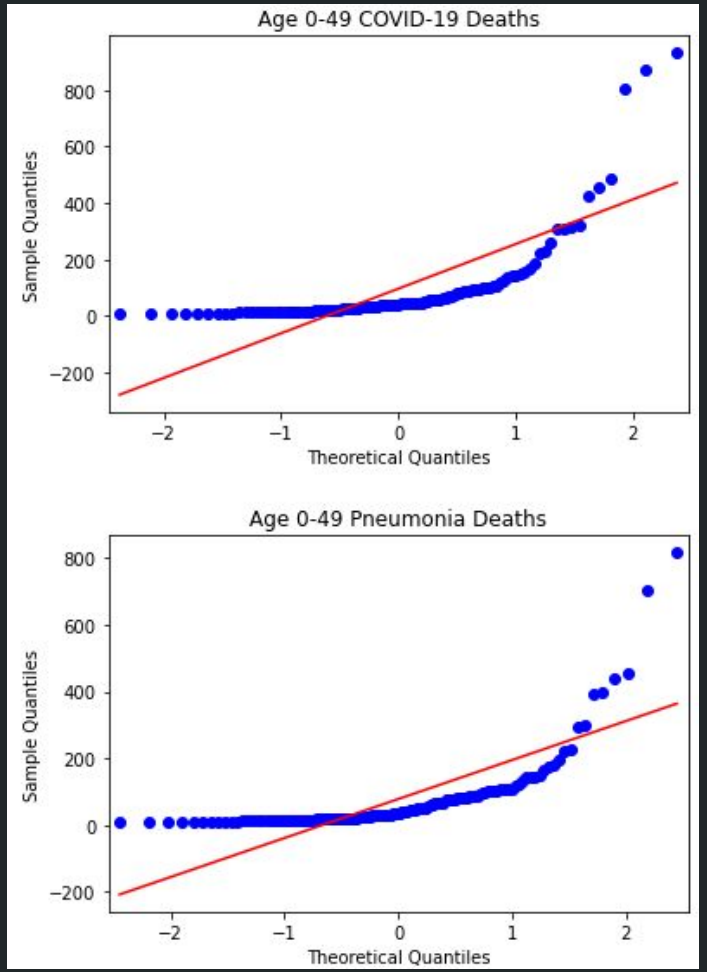


Figure 5. Each are similarly distributed

Quartile-Quartile Plots.

Red line represents normality
Both data sets are ‘non-normal’ in a similar way.

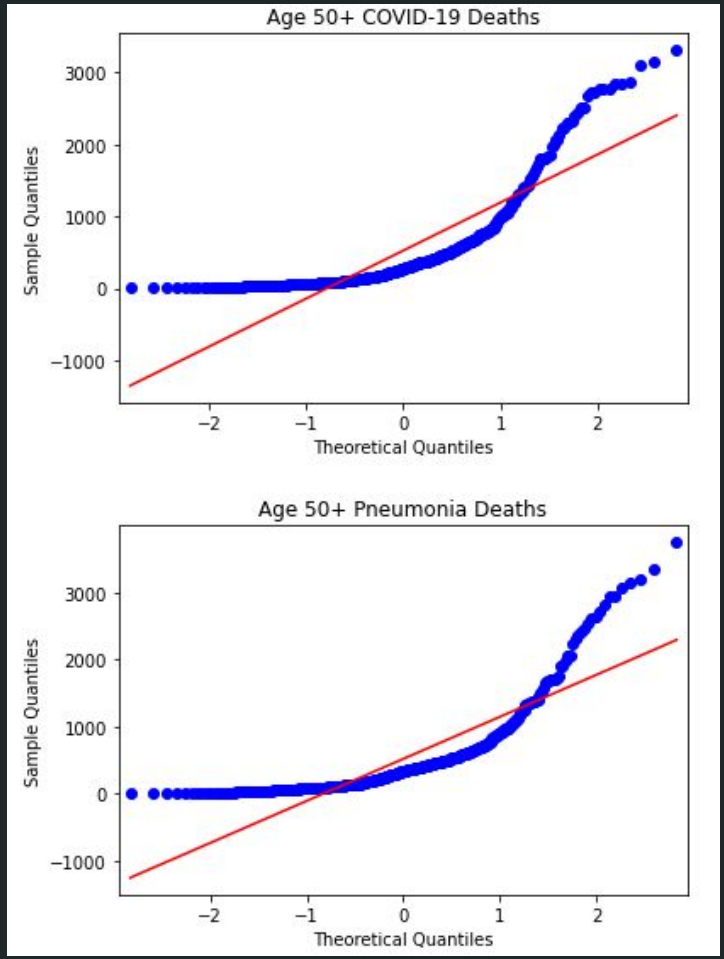
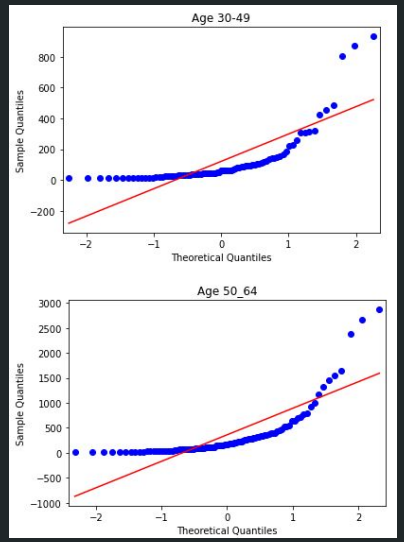
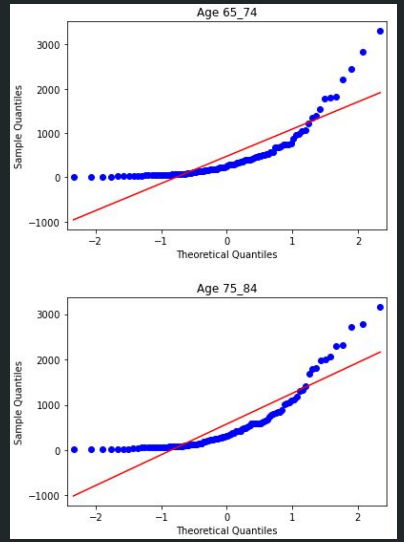


Figure 6. Each are similarly distributed

COVID-19 Quartile-Quartile Plots.

Red line represents normality
Both data sets are ‘non-normal’ in a similar way.





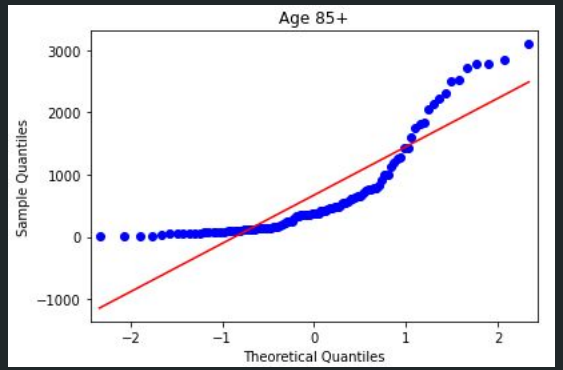
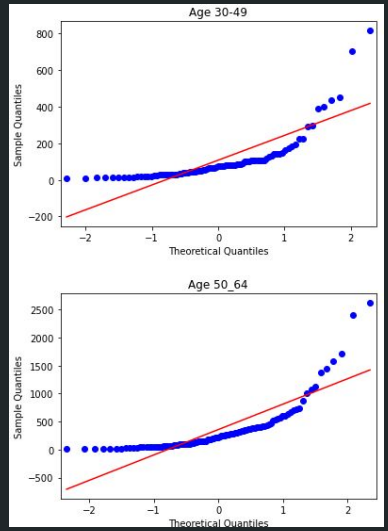
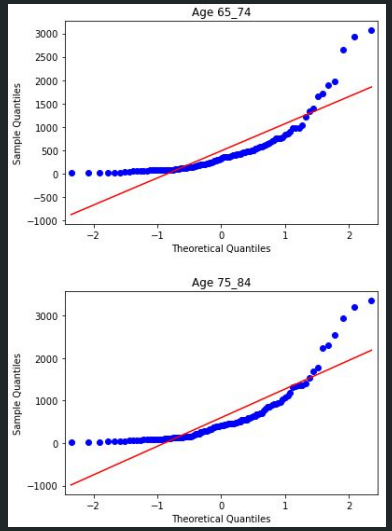


Figure 7. Each are similarly distributed

Pneumonia Quartile-Quartile Plots.

Red line represents normality
Both data sets are ‘non-normal’ in a similar way.





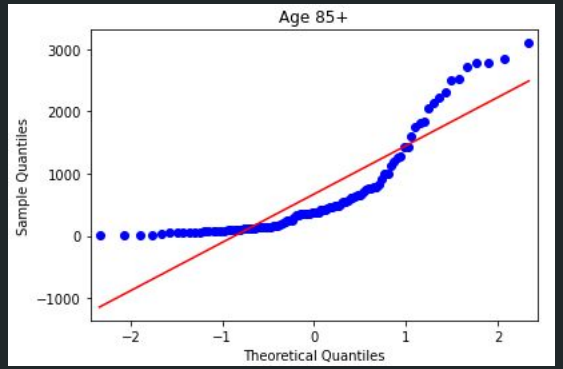


Figure 8. Each are similarly distributed

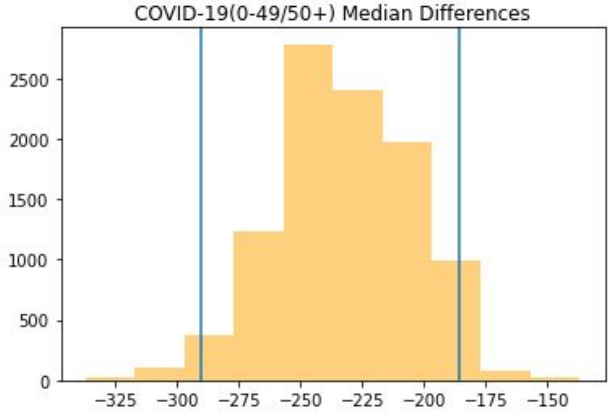
Aha!

My discoveries

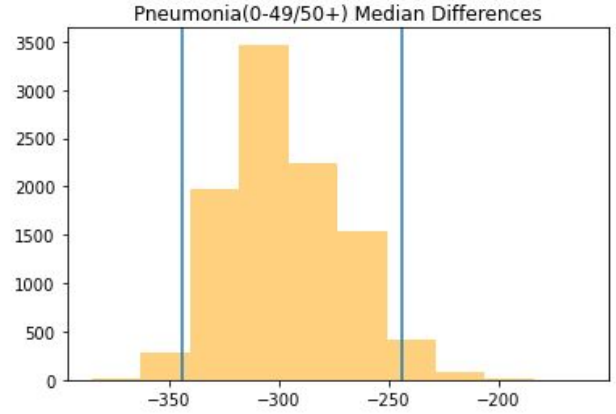
What do these confidence intervals show?

1. COVID-19: We can see from the analysis above that when considering only non-zero death counts, there are 186 to 293 more deaths reported in the '50+' age group compared to the '0-49' age group

2. Pneumonia: We can see from the analysis above that when considering only non-zero death counts, there are 244 to 344 more deaths reported in the '50+' age group compared to the '0-49' age group



(-290.0, -186.0)



(-344.0, -244.0)

Aha!

My discoveries

What about the other age groups?

This table lists the confidence interval and the common language effect size for all the age group that had a statistically significant difference in deaths reported.

For example: you can expect the 75-84 age group to have either 48-145 more deaths or about 71.5% more deaths than the 30-49 age group.

Age group combo	COVID-19	Pneumonia
(30-49 : 50-64)	(48-145, 71.5%)	(73-194, 74.6%)
(30-49 : 65-74)	(192-250, 76.7%)	(133-309, 80.8%)
(30-49 : 75-84)	(130-338, 80.0%)	(203-395, 84.7%)
(30-49 : 85+)	(163-387, 83.2%)	(223-400, 86.6%)
(50-64 : 75-84)	(27-199, 61.4%)	(47-233, 63.6%)
(50-64 : 85+)	(54-262, 65.0%)	(62-242, 64.8%)
(0-49 : 50+)	(134-264, 82.3%)	(196-319, 87.0%)

So what the doctors and medical professionals on TV says makes sense! It is important to take precautions to keep our older population safe!

Conclusion

The data used in this research shows that among records with at least one reported death there are statistically significant differences in the number of deaths attributed to COVID-19 by age group, and the older the age group the greater number of deaths. The data shows that COVID-19 and Pneumonia both display this trend.

What else can we learn from this data?

This only considers deaths. What about the survivors?

The younger generations may be able to carry the virus asymptomatically, and that could have a big impact on transmission. Is that something that can be investigated with this data set?

Are there differences between genders?

Are there other data sets that we can pair with this one?

What about deaths reported from states with a larger elderly population? Can that influence be observed?

The End

Questions?