Module - 3

ALY 6010 Probability Theory and Statistics

RAHUL AVINASH JADHAV

Northeastern University

College of Professional Studies, Northeastern University, Boston, MA 02115

Contact: jadhav.ra@northeastern.edu

Submitted to Professor: Prof. Roy Wada

Date of submission: 11/28/2021

Introduction

In this assignment we worked on covid19 dataset of Georgia state which consists of attributes as sex, ethnicity, race, cases and deaths. The dataset contained 254018 records. We learned how to perform t.test, prob test and binomial test on the dataset. We also performed bar plots, learned how to export our analysis table.

Analysis

Table 1: Different types of testing on deaths.

	method	Count	mean	statistic	p_value	conf.low	conf.high
1	Exact binomial test	254018	0.044721	11360	1.28E-11	0.043921	0.045532
2	One Sample t-test	254017	0.044721	6.63554	3.24E-11	0.043917	0.045525
3	1-sample proportions test with continuity correction	1	0.044721	46.68259	8.35E-12	0.043922	0.045534

The t-test, prop test and Binomial test were conducted against the null hypothesis that the sample mean and probability was 0.042.

Table 1 shows the results if different types of testing were performed for the death.

- When performed Binomial testing we can see that the p value was 1.28E-11 with conf. low as 0.043921 and conf. high as 0.045532
- When performed One sample t-test we can see that the p value was 3.24E-11 with conf. low as 0.043917 and conf. high as 0.045525
- When performed 1-sample proportion test we can see that the p value was 8.35E-12 with conf. low as 0.043922 and conf. high as 0.045534.

Table 2: T-test result by whole and group

		Count	mean	statistic	p_value	conf.low	conf.high
1	Whole cases	254017	0.044721	6.63554	3.24E-11	0.043917	0.045525
2	Male cases	111884	0.054225	18.05716	8.80E-73	0.052898	0.055552
3	Female cases	141778	0.037326	-9.28491	1.64E-20	0.036339	0.038312
4	Unknown	353	0.002825	-13.868	3.27E-35	-0.00273	0.008381

The t-test of the mean were conducted against the null hypothesis that the sample mean was 0.042. Table 2 displays the results of the t-tests for the positive cases of covid as whole and by group. The Female cases group mean was not statistically different from 0.042, while the Unknown group mean was found to be significantly different from 0.042 at P3.27e-35.

Bar plot 1: Proportion of cases as per gender

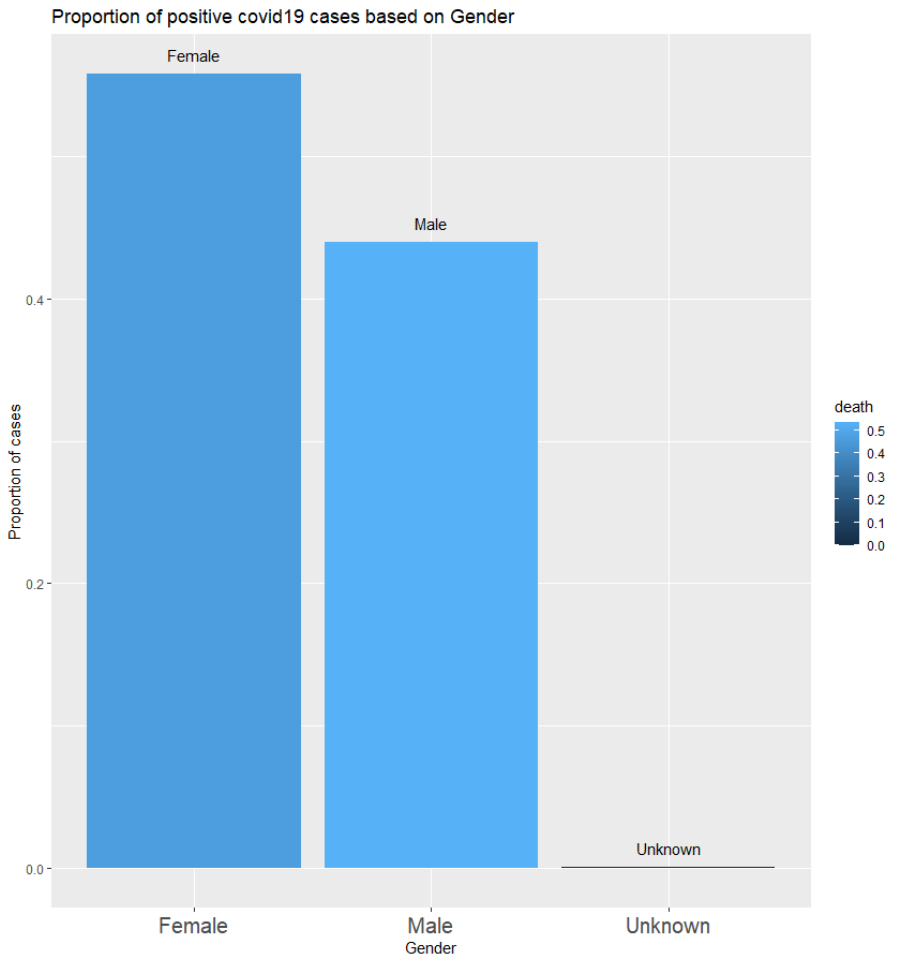


Table 3: cases and deaths based on Gender

	sex	count_cases	count_death	proportion_cases	proportion_death
1	Male	111885	6067	0.44	0.534
2	Female	141779	5292	0.558	0.466
3	Unknown	354	1	0.001	0

From bar plot 1 and table 3 we can see proportion of cases and death of people based on their gender. The number of positive cases is more in female which but if you see the death proportion, the number of deaths is more in male gender.

We can conclude that Females are more likely to contract the covid19 but the fatality is more in Male.

Bar Plot 2: Ethnicity wise Proportion of cases and death of people by chronic condition

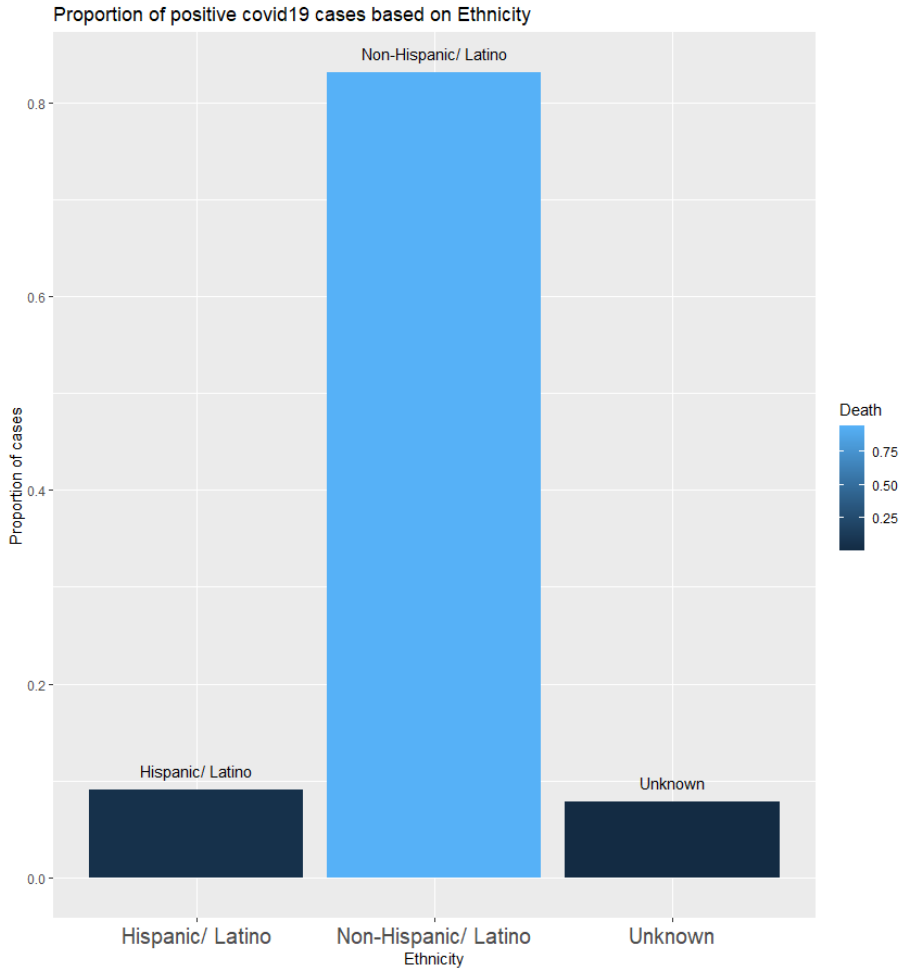
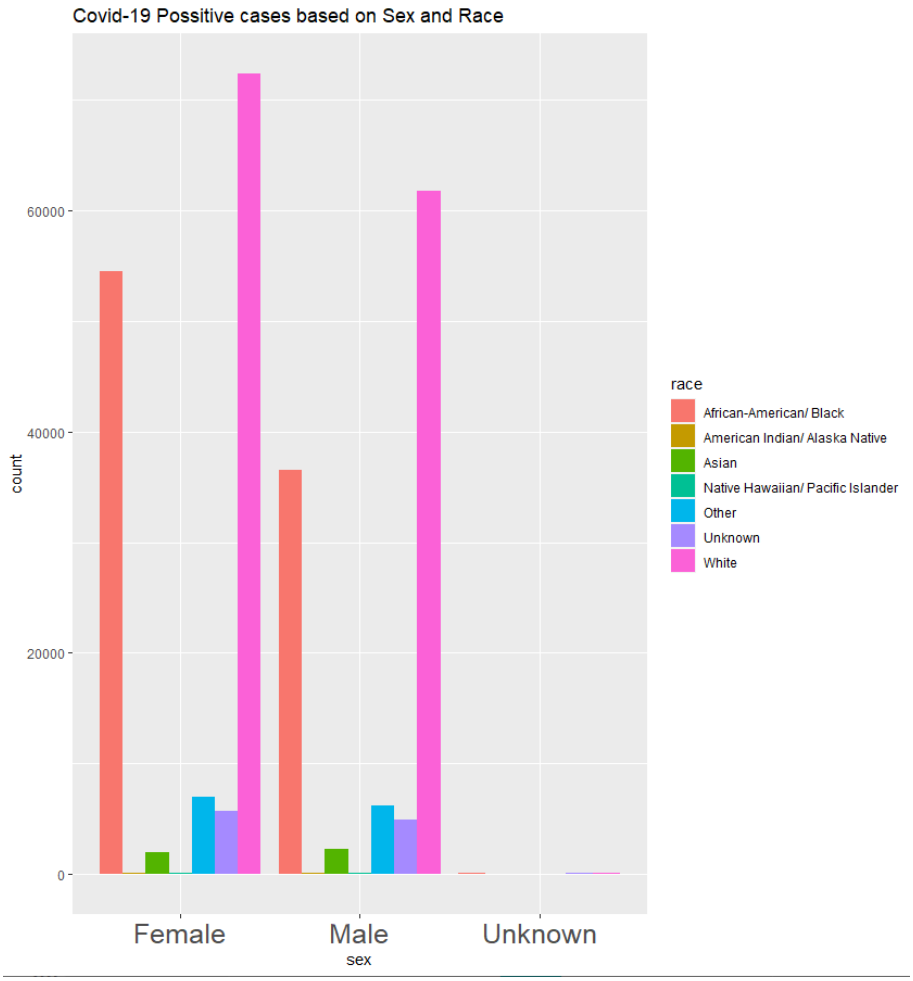


Table 4: cases and deaths based on Ethnicity

	ethnicity	count_cases	count_death	proportion_cases	proportion_death
1	Non-Hispanic/ Latino	210968	10719	0.831	0.944
2	Hispanic/ Latino	22996	601	0.091	0.053
3	Unknown	20054	40	0.079	0.004

From bar plot 2 and table 4 we can see the dataset contains more data of non-Hispanic/ Latino ethnicity. The proportion of cases and deaths are more in Non-Hispanic/Latino.

Bar Plot 3: Gender wise cases of people as per race group



From bar plot 3 we can see that Female of race white are more infected from covid. If race is to be compared, people of race white are infected more from covid followed by people of race African- American/ Black.

Summary

- From table 2 we can see that the confidence interval changes if the mean changes
- From Bar plot 1 and table 3, we can conclude that the positive cases are more in female but the death rate is more in male which concludes that females have good immunity against covid than men.
- If ethnicity is to be talked about, the records we have are more of people with ethnicity non-Hispanic/Latino.
- People of race white have more covid19 infection rate.

Bibliography

- Kabacoff, R. (2011). *R in action: Data analysis and graphics with R*. Manning.

- *Multiple condition if-else using dplyr, custom function, or purr*. (2018, August 26). Stack Overflow. https://stackoverflow.com/questions/52028764/multiple-condition-if-else-using-dplyr-custom-function-or-purr

- *R Probability*. (n.d.). Columbia Edu. http://www.columbia.edu/%7Ecjd11/charles_dimaggio/DIRE/styled-4/styled-6/code/