Abhigna Ramamurthy

NUID: 002982276

R Practice Module 3

ALY6010: Probability and introduction to statistics

October 10, 2021

Table of Contents

[Introduction 2](#_Toc84782853)

[Analyzing the comorbidity data 2](#_Toc84782854)

[Data cleaning 2](#_Toc84782855)

[Analysis 2](#_Toc84782856)

[Questions 5](#_Toc84782857)

[1. One-sample t-tests 5](#_Toc84782858)

[2. Multiple one-sample t-tests of proportion 8](#_Toc84782859)

[Conclusion 10](#_Toc84782860)

[References 10](#_Toc84782861)

# Introduction

A hypothesis testing helps in decision making process and evaluate population impact or claims based on sample data. Such testing is extremely helpful as in most real-life cases we do not have access to quantify the data for the entire population (Bluman, 2019). Data used in this report has individual data about comorbidity with values of sex, ethnicity and race along with case and death columns. This report outlines the analysis of the spread of the data along with hypothesis using R.

# Analyzing the comorbidity data

## Data cleaning

Column of comorbidity was removed as it had no significance for analysis. The rows which had unknown data under sex, ethnicity and race (all 3 unknowns) were removed. Total of 122 rows were removed as a result. Rest of the Unknown values were kept as at least one of the columns (sex, race, ethnicity) were present.

## Analysis

Understanding the dataset given –

Chart, bar chart

Description automatically generated

The above Figure 1 shows the spread of data with respect to sex. We can see that majority of the data is of female. Unknown count is very small with only 232 entries.

Chart, bar chart

Description automatically generated

Figure 2 above shows the spread of data by ethnicity. We can see that majority of them are Non-Hispanic/Latino. There is also a considerable number of entries for unknown ethnicity.

Chart, bar chart

Description automatically generated

Figure 3 above shows the distribution of data based on Race. We can see that the highest number of cases came from White and African American/Black communities.

Chart, bar chart

Description automatically generated

The figure 4 above shows the number of cases reported and number of deaths from those cases. Total number of cases is 253896. Total number of deaths is 11359. We can see that the death probability is 4.473879%.

Questions:

1. One-sample t-tests  
   Calculate the probability of death (i.e., COVID-19 infection fatality rate) and conduct one-sample t-tests of mean or proportion using a hypothesized value (usually 0 or the national norm). Supposed the true population COVID-19 infection fatality rate during the same period was 0.042. Use prop.test() or binom.test() for testing the proportion in R. Was the infection fatality rate in Georgia statistically different from the hypothesized population rate during this period? Repeat the test using t.test( ), which conducts the test of means. This can be done because probability of death is the mean of a binary variable death. Do you obtain larger or smaller p-values? Which of the three t-test commands (prop.test( ), binom.test( ), and t.test( )) appear to be the most reliable?   
     
   **Prop Test:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **estimate** | **statistic** | **p.value** | **parameter** | **conf.low** | **conf.high** | **method** | **alternative** |
| 0.044738791 | 210492.7517 | 0 | 1 | 0.043941531 | 0.045549827 | 1-sample proportions test without continuity correction | two.sided |

Table 1: Prop test for proportion of death in the cases reported from the sample given

The above is the result of prop.test() for the deaths data from the given comorbidity dataset. For a 95% confidence interval, we get the p value to be 0.04473879, which is equal to the probability of death we found in the previous step. We also get the true probability or death in the population to be between 0.04394153 and 0.04554983. The true population fatality rate is given as 0.042 which is outside the confidence interval values. Hence, we can conclude that Georgia in fact had a different fatality rate compared to the hypothesized population fatality rate.   
  
**Binomial Test:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **estimate** | **statistic** | **p.value** | **parameter** | **conf.low** | **conf.high** | **method** | **alternative** |
| 0.044738791 | 11359 | 9.59891E-12 | 253896 | 0.043937897 | 0.045550097 | Exact binomial test | two.sided |

Table 2: Binomial test for proportion of death in the cases reported from the sample given

Above screenshots show the binomial test conducted for the same set of values. The results shows that the probability of death is 0.04473879 for the given sample. With a 95% confidence interval, we can determine that the population probability will be in the interval of 0.0439379 and 0.0455501. The confidence interval values using binomial tets method has moved further away from the population fatality rate.

A picture containing graphical user interface

Description automatically generated  
The above graph shows the pictorial representation of the binomial test done on the given sample. The values input are as follows: number of successes(x) i.e., number of deaths is equal to 11359, total number of trials(n) is equal to total number of cases in the given sample is 253896, the probability of success(death) for the given sample is 0.04473879. The x-axis shows the number of deaths seen in this sample and the y axis shows the probability of death in a sample of 253896 trails.

**T Test:**

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| **estimate** | **statistic** | **p.value** | **parameter** | **conf.low** | **conf.high** | **method** | **alternative** |
| 0.044738791 | 109.0455942 | 0 | 253895 | 0.043934661 | 0.045542921 | One Sample t-test | two.sided |

Table 3: T test for proportion of death in the cases reported from the sample given

In the above screenshots, a t-test has been performed on the death column of the data provided. The mean value of the deaths is 0.04473879. With a 95% confidence interval, the t-test gives the probability of death to be between 0.04393466 and 0.04554292. This has moved away from the population fatality rate but has a smaller interval when compared to binomial test results.   
  
Based on the analysis made above, proportion test confidence interval results seem to be more accurate and closer to the population fatality rate.

1. Multiple one-sample t-tests of proportion  
   Conduct multiple one-sample t-tests of proportion (prop.test or binom.test) for multiple groups. Note that you may need to create two vectors to isolate a group. For example, conducting a t-test for a male sample and another t-test for a female sample requires filter () into two vectors, and then run a test for each sample. Are they different from the hypothesized population rate (0.042)? Make decision based on p-value. Do this in R. Provide the null and the alternative hypothesis for each test. Provide the test results in a table and several paragraphs interpreting the results.

Prop test for male and female groups:

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **estimate** | **statistic** | **p.value** | **parameter** | **conf.low** | **conf.high** | **method** | **alternative** |
| Female | 0.037326 | 121401.1 | 0 | 1 | 0.036351 | 0.038325 | 1-sample proportions test without continuity correction | two.sided |
| Male | 0.054225 | 88932.94 | 0 | 1 | 0.052914 | 0.055568 | 1-sample proportions test without continuity correction | two.sided |

Table 4: Multiple one-sample prop test for deaths in female and male groups from the given sample

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **estimate1** | **estimate2** | **statistic** | **p.value** | **parameter** | **conf.low** | **conf.high** | **method** | **alternative** |
| 0.037325697 | 0.054225321 | 417.5362631 | 1 | 1 | -0.018287369 | 1 | 2-sample test for equality of proportions without continuity correction | greater |

Table 5: Two-sample prop test for deaths in female and male groups from the given sample.

Null hypothesis for this testing is fatality rate of female is lower than that of male population. Negative confidence interval low and high values represent that the fatality rate for female is lower than that of male population. We can also see that the estimates for female fatality rate is 0.03 whereas for male is 0.05. Hence, we can conclude that the fatality rate of female population is lower than that of the male population and accept the null hypothesis.   
  
Prop test for ethnicity

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
|  | **estimate** | **statistic** | **p.value** | **parameter** | **conf.low** | **conf.high** | **method** | **alternative** |
| Hispanic/ Latino | 0.02613498 | 20654.82849 | 0 | 1 | 0.024150809 | 0.028277442 | 1-sample proportions test without continuity correction | two.sided |
| Non-Hispanic/ Latino | 0.050808653 | 170270.4718 | 0 | 1 | 0.049879706 | 0.051753959 | 1-sample proportions test without continuity correction | two.sided |
| Unknown | 0.001956653 | 19776.30524 | 0 | 1 | 0.001431733 | 0.002673509 | 1-sample proportions test without continuity correction | two.sided |

Table 6: Multiple one-sample prop test for deaths in different ethnicities from the given sample.

|  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- |
| **estimate1(Hispanic/ Latino)** | **estimate2(Non-Hispanic/ Latino)** | **statistic** | **p.value** | **parameter** | **conf.low** | **conf.high** | **method** | **alternative** |
| 0.02613498 | 0.050808653 | 274.1751729 | 1 | 1 | -0.026574457 | 1 | 2-sample test for equality of proportions without continuity correction | greater |

Table 7: Two-sample prop test to compare the deaths proportion in Hispanic/Latino and Non-Hispanic/Latino ethnicities.

Null hypothesis in this test is highest number of cases were reported from Non-Hispanic/Latino ethnic community. The unknown ethnicity values are not considered here as their estimated value from Table 6 was way smaller than that of the other two ethnicities. We can see that the estimate for Hispanic community is lesser than that of the non-Hispanic community. The negative lower value of interval also shows that the Hispanic community fatality rate is lower than that of the non-Hispanic ethnicity and hence the null hypothesis is true and can be accepted.

# Conclusion

The sample data give is the for patients with multiple chronic disorders and are reported positive for covid virus. The cases who had unfortunate incident of death are also recorded. This data can be aggregated to make hypothesis based on the value. Questions like which gender has higher probability of death, which ethnicity has higher probability of death are explored. This also helps in understanding how the hypothesis testing works. Further analysis can include data like age group, geographical location and also what were the actual chronic condition.

# References

Cross Validated. 2021. Multiple one-sample t Test. [online] Available at: <https://stats.stackexchange.com/questions/441016/multiple-one-sample-t-test> [Accessed 10 October 2021].

DataScience Made Simple. 2021. *Delete or Drop rows in R with conditions - DataScience Made Simple*. [online] Available at: <https://www.datasciencemadesimple.com/delete-or-drop-rows-in-r-with-conditions-2/> [Accessed 10 October 2021].

Datatofish.com. 2021. *How to Create DataFrame in R (with Examples) - Data to Fish*. [online] Available at: <https://datatofish.com/create-dataframe-in-r/> [Accessed 10 October 2021].

Phillips, N., 2021. *YaRrr! The Pirate’s Guide to R*. [online] Bookdown.org. Available at: <https://bookdown.org/ndphillips/YaRrr/creating-matrices-and-dataframes.html> [Accessed 10 October 2021].

Rdocumentation.org. 2021. *binom.test function - RDocumentation*. [online] Available at: <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/binom.test> [Accessed 10 October 2021].

Rdocumentation.org. 2021. *binom.test function - RDocumentation*. [online] Available at: <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/binom.test> [Accessed 10 October 2021].

Rdocumentation.org. 2021. *prop.test function - RDocumentation*. [online] Available at: <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/prop.test> [Accessed 10 October 2021].

Rdocumentation.org. 2021. *t.test function - RDocumentation*. [online] Available at: <https://www.rdocumentation.org/packages/stats/versions/3.6.2/topics/t.test> [Accessed 10 October 2021].

Rdrr.io. 2021. *ggproptest: Plot test of Equal or Given Proportions in gginference: Visualise the Results of Inferential Statistics using 'ggplot2'*. [online] Available at: <https://rdrr.io/cran/gginference/man/ggproptest.html> [Accessed 10 October 2021].

Stack Overflow. 2021. *Change size of axes title and labels in ggplot2*. [online] Available at: <https://stackoverflow.com/questions/14942681/change-size-of-axes-title-and-labels-in-ggplot2> [Accessed 10 October 2021].

Statmethods.net. 2021. *Quick-R: Bar Plots*. [online] Available at: <https://www.statmethods.net/graphs/bar.html> [Accessed 10 October 2021].

Sthda.com. 2021. *ggplot2 barplots : Quick start guide - R software and data visualization - Easy Guides - Wiki - STHDA*. [online] Available at: <http://www.sthda.com/english/wiki/ggplot2-barplots-quick-start-guide-r-software-and-data-visualization> [Accessed 10 October 2021].

Sthda.com. 2021. *ggplot2 title : main, axis and legend titles - Easy Guides - Wiki - STHDA*. [online] Available at: <http://www.sthda.com/english/wiki/ggplot2-title-main-axis-and-legend-titles> [Accessed 10 October 2021].

Sthda.com. 2021. *One-Proportion Z-Test in R - Easy Guides - Wiki - STHDA*. [online] Available at: <http://www.sthda.com/english/wiki/one-proportion-z-test-in-r> [Accessed 10 October 2021].

Youtube.com. 2021. [online] Available at: <https://www.youtube.com/watch?v=-msRQ0YZtAY> [Accessed 10 October 2021].

Zach, 2021. *How to Plot a Binomial Distribution in R - Statology*. [online] Statology. Available at: <https://www.statology.org/plot-binomial-distribution-r/> [Accessed 10 October 2021].