Statistical Analysis with Bootstrap Sampling

COVID-19 Cases in Toronto

Abrar Nasir

3/10/2021

Part 2

Introduction

Sourced from the City of Toronto's Open Data Portal, an exploratory data analysis on COVID-19 cases in Toronto is pursued as the literature follows. The core methodology used in this report is bootstrapping which will help the reader to understand what proportion of confirmed patients were female as well as the mean age of confirmed patients. The aforementioned method is used to estimate finite approximations of the parameters since the raw sample is fairly small and the distributions not normal.

The analysis of COVID-19 cases data is extremely employable in medical, economic and socio-cultural fields all over the globe as it can help individuals comprehend the trends associated with the pandemic which will, in turn, allow people to make more informed decisions.

A hypothesis can be proposed in this report, particularly in the analysis of the relation between patient age and the number of confirmed COVID-19 patients, such that adults around the age of 42-44. For the gender distribution of COVID-19 patients, it can be assumed more number of females were affected based on the fact that there are more females in Toronto than males. (T)

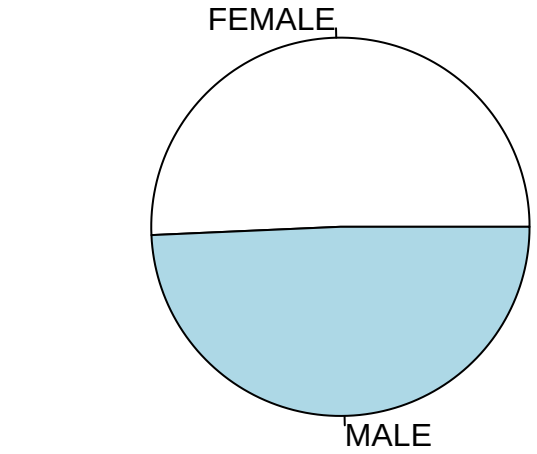
Data

The data has been imported on our statistical software (R version 4.0.4 - Lost Library Book) using a helper package, namely *opendatatoronto* and the source is managed by Toronto Public Health since January 2020 and updates are pushed every week. (Open Data Toronto)

The dataset had to be cleaned to get rid of irrelevant variables such as source of infection, neighborhood name, assigned ID to name a few. Only rows containing information of confirmed COVID-19 patients and columns of Age Group and Client Gender were maintained. However, the Age Group column had to be replaced with a column which stored the midpoint values of the age classes. Moreover, only for the simplification of this analysis, entries with gender other than male and female were removed. (Suhani, Kassambara) (Stack Overflow) (R/Rstats)

The table below shows the centre and spread statistics for the age data from the raw sample data:

	Min	1st Quartile	Median	Mean	Std. Deviation	3rd Quartile	Max
Age	10.00	24.50	44.50	42.87	22.2159	54.50	99.00



The visualization above also shows the proportions of male and female patients as per the COVID-19 cases dataset, from where we can tell there were slightly more females who were affected by the pandemic.

Methods

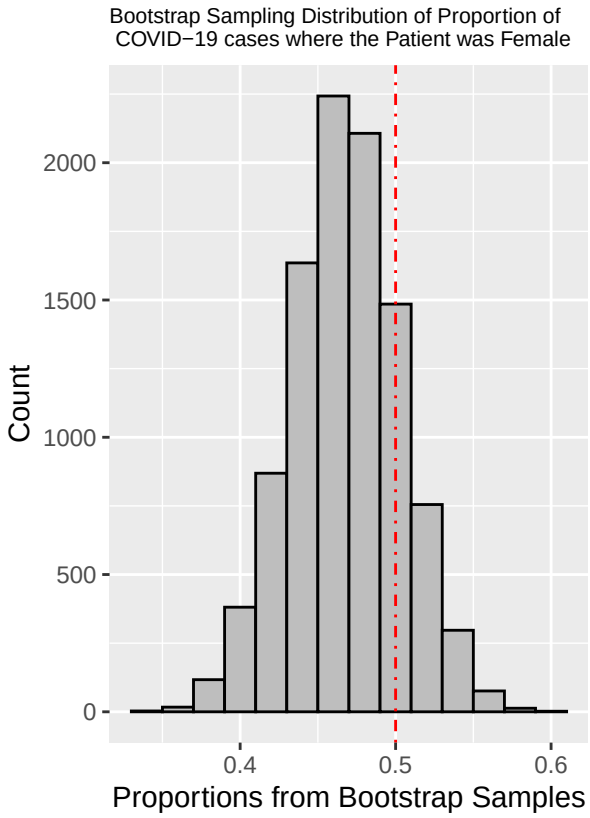
Bootstrapping, in layman terms, is essentially the statistical procedure of resampling our given dataset to create multiple simulated samples, to calculate truer estimates of confidence intervals and perform hypothesis testing for sample statistics. (Frost, J) For this analysis, we are to focus on empirical bootstrap over parametric bootstrap owing to the fact that our analysis questions do not possess any distribution which is predefined by a known parameter. Moreover, we have elements in the data which can be used to find the means and proportion sampling distribution with confidence intervals, which influences our choice of bootstrap scheme.

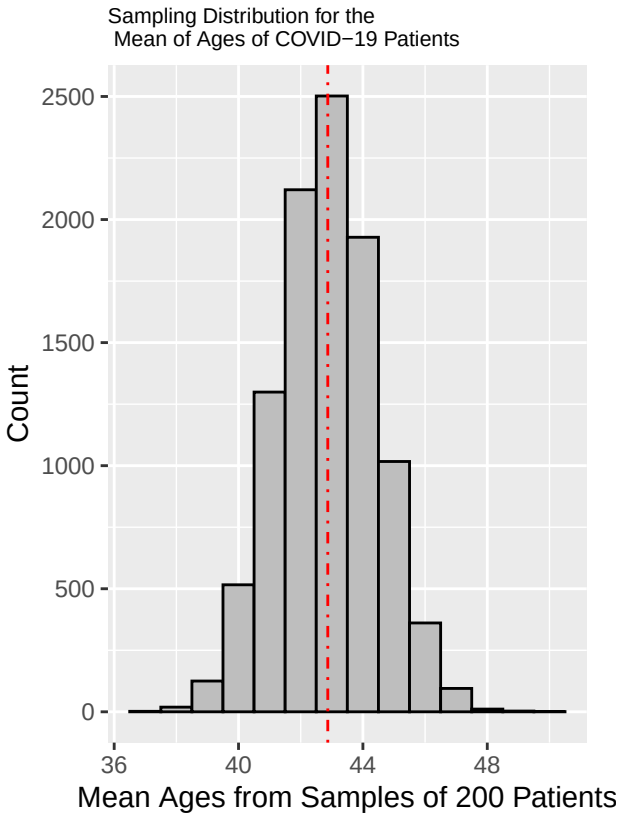
For both our analysis of age, a 90% confidence interval will be assessed to explore mean as the parameter of interest. On the other hand, for gender, proportion of females from the confirmed COVID-19 patients list will be analysed, with information about the 90% confidence interval as well.

For both cases, sample sizes of 200 entries are used while 10,000 replications are done to find the bootstrap samples.

For this part of the analysis, R version 4.0.4 - Lost Library Book has been used as well.

Results





90% Confidence Intervals (CI):

	5%	95%
Proportion of Female COVID-19 patients	0.41	0.53
Mean Age of COVID-19 patients	40.32	45.46

The histograms, closely resembling distributions of a Gaussian function, are visualisations showing the bootstrap sampling distribution of the proportion of COVID-19 cases where the patient was female and bootstrap sampling distribution of the mean age of COVID-19 patients.

From the CI table, we can tell that there is a 90% chance that, from a given population of COVID-19 cases in Toronto, we will get 41% to 53% of the patients to be female. However, as per 2016 census, provided that there were more men than women in Toronto, this scenario slightly disagrees with the stated hypothesis where it was assumed more female will be affected by COVID-19 and the disagreement can be understood by observing the red vertical line at 50% line to towards the right of the longest bin of the histogram.

In case of mean age of confirmed COVID-19 patients, it can inferred that 90% of the population who got sick due to COVID-19 are of age ranging between 40.32 and 45.46. This statement positively reacts to the hypothesis made earlier and it can also be seen that the mean age of highest probability found through bootstrapping is somewhat around the mean age of the raw sample data that has been used in this analysis.

Conclusions

Finally, we can tell that our bootstrapped samples outputted histograms showing similarities to a normal distribution, which shows the reliability of our testing. For proportions of female COVID-19 patients, the 90% CI was between 41% and 53% which is reasonable which that for mean age of COVID-19 patients have a 90% possibility of being between 40.32 and 45.46 which is quite logical as well, and agrees with the hypothesis which said the mean should be around 45 years old.

A better, reliable analysis could have done if the bootstrap replication was repeated more than 10,000 times with larger sample sizes. However, due to limitation in time and computing power and efficiency, the trade-off had to be made but in future reports, it can be said for sure higher levels of simulations will be pursued.

Bibliography

1. Open Data Toronto. (n.d.). Retrieved March 11, 2021, from https://open.toronto.ca/dataset/covid-19-cases-in-toronto/

2. Frost, J. (2020, June 12). Introduction to bootstrapping in statistics with an example. Retrieved March 11, 2021, from https://statisticsbyjim.com/hypothesis-testing/bootstrapping/

3. Suhani, Kassambara. (2018, October 19). Rename data frame columns in r. Retrieved March 11, 2021, from https://www.datanovia.com/en/lessons/rename-data-frame-columns-in-r/

4. Stack Overflow (n.d.). How to get midpoint from range in string within a dataframe column in r? Retrieved March 11, 2021, from https://stackoverflow.com/questions/66565888/how-to-get-midpoint-from-range-in-string-within-a-dataframe-column-in-r

5. T. (n.d.). POPULATION DEMOGRAPHICS. Retrieved March 10, 2021, from https://www.toronto.ca/wp-content/uploads/2019/11/99b4-TOHealthCheck_2019Chapter1.pdf

6. R/Rstats - create new column of data based on criteria? (n.d.). Retrieved March 11, 2021, from https://www.reddit.com/r/rstats/comments/7pgc1x/create_new_column_of_data_based_on_criteria/