Investigating the COVID-19 Outbreak

Abby Mapes, Delaney Demark, Jenny Huang, Harsha Srijay

Packages

Load and Clean Data

Introduction

Data Analysis Plan

Through our analysis, we will use death as our outcome variable by analyzing the proportion of patients who have died, indicated by a value of "yes" for death. To do so, we will use the following predictor variables: age_group, gender, visited_wuhan, from_wuhan, country.

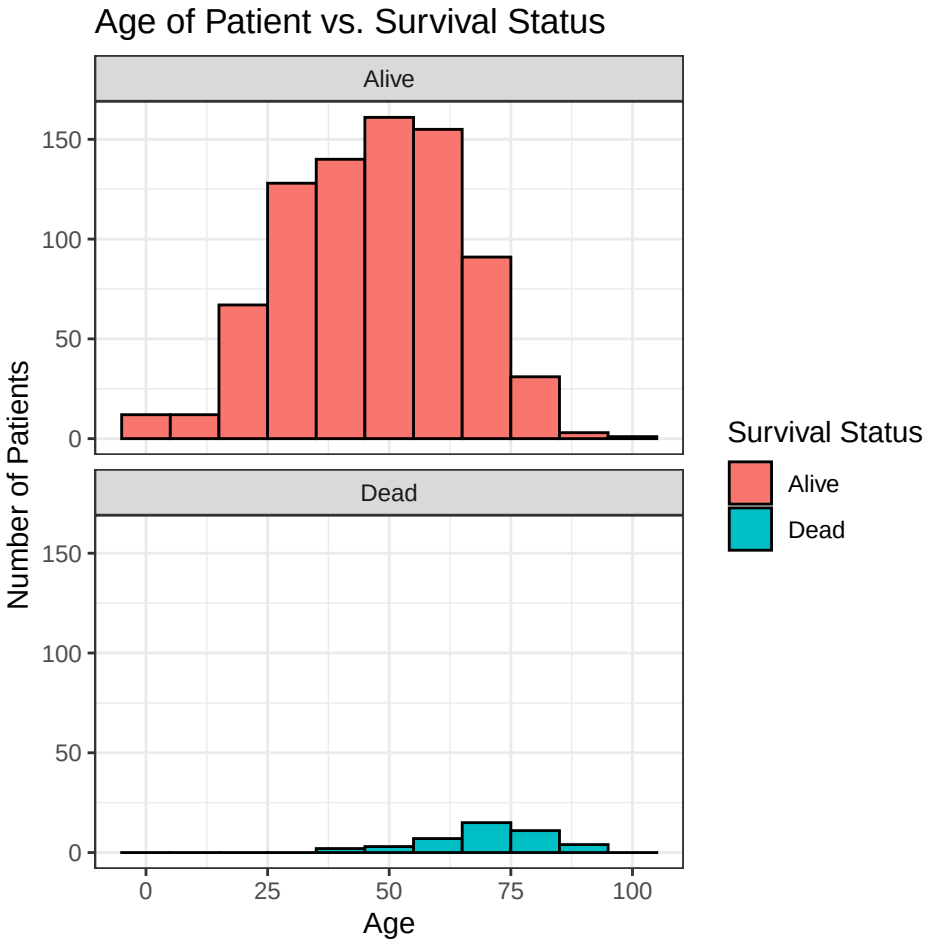
Using these variables, we will attempt to determine not only if age affects one's survival outcome due to COVID-19, but also if any of these other characteristics are associated with one's survival outcome.

To start, we will preform some preliminary exploratory data analysis to learn more about our data.

First, we will determine the death rate, proportion of those dead, of all patients in our dataset. As we can see, there is a small percentage, about 4%, of patients who died from COVID-19 in our data set.

```
# A tibble: 2 x 3
  death     n prop_dead
  <chr> <int>     <dbl>
1 Alive  1043    0.961
2 Dead     42    0.0387
```

Now, we will visualize the ages of patients for each survival status.



Below is a summary table of the mean age of patients who have died from COVID-19 and those that haven't, out of all patients in our data set that reported their ages. As we can see, the average age of patients that have died is greater than the average age of patients that are alive. These ages give us a reference point to determine what age will be considered "old" and what ages will be considered "young" for our exploratory data analysis.

```
# A tibble: 2 x 2
  death mean_age
  <chr>    <dbl>
1 Dead      70.1
2 Alive     48.4
```

To understand some of our other explanatory variables, we will calculate some statistics to get a sense of our data in terms of gender, country, and patients who have been to Wuhan recently. As we can see below, around 35% of patients are femalee, 48% of patients are male and 17% of patients are not classified. Noting that more male patients are included in our data set will be important and helpful when performing our exploratory data analysis.

```
# A tibble: 3 x 3
  gender     n  prop
  <chr>  <int> <dbl>
1 female   382 0.352
2 male     520 0.479
3 <NA>     183 0.169
```

Additionally, the patients in our data set come from 38 different countries.

```
# A tibble: 1 x 1
  total_countries
            <int>
1              38
```

However, as we can see below, there are only 3 countries with patients who have died in our data set: China, Hong Kong, Taiwan. It will be helpful to know that only 3 of the 38 total countries in our data set have reported deaths when we consider the explanatory variable country in our exploratory data analysis.

```
# A tibble: 3 x 3
# Groups:   country [3]
  country       n prop_dead
  <chr>     <int>     <dbl>
1 China        39    0.198
2 Hong Kong     2    0.0213
3 Taiwan        1    0.0294
```

Additionally, from the summary table below, we see the mean age of the patients who have died for each country where there are reported deaths. In China, the mean age is about 71. In Taiwan, the mean age is 65. In Hong Kong, the mean age is about 54.

```
# A tibble: 3 x 2
  country   mean_death_age
  <chr>              <dbl>
1 China               71.1
2 Taiwan              65
3 Hong Kong           54.5
```

From the table below, we see the majority of patients included in our data set are not from Wuhan, nor have they reported that they have previously visited Wuhan. Even though the majority of patients in our data set are not from or have been to Wuhan, it will be interesting to see if time in Wuhan is associated with one's survival outcome.

```
# A tibble: 3 x 2
  from_wuhan    prop
  <chr>        <dbl>
1 No         0.853
2 Yes        0.144
3 <NA>       0.00369
```

```
# A tibble: 2 x 2
  visiting_wuhan  prop
  <chr>          <dbl>
1 No             0.823
2 Yes            0.177
```

In our exploratory data analysis, we plan to use the following statistical methods to answer some of these questions:

a) Calculate the conditional probability of P(Death | Age) and determine if there are any confounding variables: gender, visted_wuhan, from_wuhan, country.

b) Take a bootstrap sample to estimate a confidence interval for the mean age of affected individuals. If there are any confounding variables, we will estimate the mean age of affected individuals faceted by the confounding variables.

c) Run a hypothesis test with significance level of .05 to assess claims about potential associations between death rate and our explanatory variables. Death Rate

- H0: The death rate for older individuals is less or equal to the death rate for younger individuals.

- H1: The death rate for older individuals is significantly higher than younger individuals.

  After running the hypothesis test, if we get a p-value $< .05$, then we will be able to reject that the death rate for older individuals is less than or equal to the death rate for younger individuals.

Wuhan - H0: The death rate for those who have visited or lived in Wuhan is the same as the death rate for those that haven't. - H1: The death rate for those who have visited or lived in Wuhan is significantly different than those that haven't.

```
After running the hypothesis test, if we get a p-value < .05, then we
will be able to reject that the death rate is the same for patients who
have visited or lived in Wuhan as those that haven't.
```

Gender - H0: The death rate for women is the same as the death rate for men. - H1: The death rate for women is different than the death rate for men.

```
After running the hypothesis test, if we get a p-value < .05, then we
will be able to reject that the death rate is the same for women patients
as it is for male patients.
```

d) Conduct a permutation test with significance level of .05 to determine if death rate and age are independent variables.
   - H_0: Death rate and age group are independent.
   - H_1: Death rate and age group are not independent. In fact, the death rate of older individuals is higher than the death rate of younger individuals.
     After running the hypothesis test, if we get a p-value $< .05$, then we will be able to reject that the death rate and age group are independent.

Data

```
Observations: 1,085
Variables: 17
$ id                    <fct> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14...
$ reporting_date        <date> 2020-01-20, 2020-01-20, 2020-01-21, 2020-01-...
$ location              <chr> "Shenzhen, Guangdong", "Shanghai", "Zhejiang"...
$ country               <chr> "China", "China", "China", "China", "China", ...
$ gender                <chr> "male", "female", "male", "female", "male", "...
$ age                   <dbl> 66, 56, 46, 60, 58, 44, 34, 37, 39, 56, 18, 3...
$ symptom_onset         <date> 20-01-03, 2020-01-15, 20-01-04, NA, NA, 2020...
$ if_onset_approximated <chr> "No", "No", "No", NA, NA, "No", "No", "No", "...
$ hosp_visit_date       <date> 20-01-11, 2020-01-15, 2020-01-17, 2020-01-19...
$ exposure_start        <date> 2019-12-29, NA, NA, NA, NA, NA, NA, 20-01-10...
$ exposure_end          <date> 20-01-04, 20-01-12, 20-01-03, NA, NA, NA, NA...
$ visiting_wuhan        <chr> "Yes", "No", "No", "Yes", "No", "No", "No", "...
$ from_wuhan            <chr> "No", "Yes", "Yes", "No", "No", "Yes", "Yes",...
$ death                 <chr> "No", "No", "No", "No", "No", "No", "No", "No...
$ recovered             <chr> "No", "No", "No", "No", "No", "No", "No", "No...
$ source                <chr> "Shenzhen Municipal Health Commission", "Offi...
$ age_group             <chr> "60-79", "40-59", "40-59", "60-79", "40-59", ...
```