# 讨论:ISBD Applied Statistics 2020Spring Project Proposal 03

## Motivation

March 11, 2020, the World Health Organization (WHO) declared COVID-19 a global pandemic. COVID-19 is an infectious disease caused by severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). It has now (time up to 0:00, April 6, Beijing time) spread over 208 countries or territories with 1,136,851 reported cases, claiming the lives of 62,955. Amid the escalating fear over the spread of the disease, it is nonetheless often reported in the mass media that vulnerability to COVID-19 is highly age-specific, with older adults the most vulnerable to its worst effects. Indeed, according to a recent study conducted by the China Centers for Disease Control and Prevention (China CDC), the case fatality rate (CFR; proportion of deaths among the confirmed cases) for patients 70 years or older can be as high as 8%-14.8%, while that for the non-elderly (<50 years old) remains steadily below 0.4%.

To confirm the validity of such reports and findings, we aim to formally assess the age difference in the case fatality rate by analyzing publicly available COVID-19 epidemiological datasets using survival analysis methodology.

With the deepening of the analysis, we found that gender has a great relationship with the mortality of different age groups. At a further level, we want to find a suitable description of the impact of confirmed time on the case fatality rate.

目录



## Data Set

To that end, we use the online data pulled from the Korea Centers for Disease Control & Prevention (Korea CDC) and prepared by the DS4C (Data Science for COVID-19) Project. The particular dataset of interest is `PatientInfo.csv`, which contains subject-level data for over 2000 confirmed COVID-19 cases in South Korea.

First import the libraries needed in following analysis.

```
library("survival")
library("survminer")
data('lung')
library('coin')
library('party')
```

Then load the datasets `PatientInfo.csv`, and have a glimpse at it.

```
a <- read_csv("PatientInfo.csv", na = "NA")
glimpse(a)
```

Key variables include:

```
patient_id: the ID of the patient
sex: the sex of the patient
age: the age of the patient
confirmed_date: the date of being confirmed
released_date: the date of being released
deceased_date: the date of being deceased (death)
state: isolated / released / deceased
```

# Ideal Data Cleaning

To perform the data cleaning, I use different methods for different patients. My main ideas are as follows.

## Missing State

For those with a missing state variable, I will just drop them. That was because from my perspective, state is an important variable in survival analysis. Without state, we do not know whether an observation is deceased or not. To avoid any mistake caused by incorrect state imputation, I choose to drop those with a missing state variable.

## Isolated Patients

For the isolated patients, I will use the date on which the data set is last updated (2020-03-21) as the censoring date.

## Released Patients

For the released patients, I will use the release date as censoring date. If the released date is not provided but the confirmed date is provided, I will use the “confirmed date + overall average isolation time duration (which is about 14 days)” as the censoring date.

## Decreased Patients

For the deceased patients, the deceased date is just the death date. For those with a missing deceased date, if the confirmed date is provided, I will use “confirmed date + overall average time duration between confirmed and death” as the death date.

## Missing Confirmed Data

For those with a missing confirmed date, if the censoring/death date is available from above methods and the original data set, I will perform a weighted sampling to pick a date between the first confirmed date among all the patients and this patient’s censoring/death date as his/her confirmed date. The weight is generated from the counts of every confirmed dates: The more patients were confirmed during a specific date, the higher weight this date will receive.

## Missing Age

For those with a missing age, if the birth year is available, I will calculate his/her age from his/her birth year. If a patient has both missing age and missing birth year, I will drop this observation because of the importance of age in this analysis.

## Missing Gender

For those with a missing gender, I will randomly assign a gender to them: the probabilities of being assigned to be a male and being assigned to be a female are the same.

## Outcome of interest

Time from being confirmed and censoring/death is the outcome of interest. For those with a confirmed date after his/her deceased date, I will drop them because there is a high possibility that the recording is wrong.

But actually some data has unexpected symbols, so we can't distinguish between erroneous data and missing data. In fact, we only used a part of the data without problems.

```
a[1:794,]
age = as.numeric(unlist(strsplit(as.character(a$age), split = 's')))
D = data.frame(date = a$confirmed_date, status = rep(1, nrow(a)))

D$date = as.Date(as.character(D$date))
D$date = D$date - D$date[1]
D$age = cut(age, breaks = c(0, 40, 50, 60, 70, 80, Inf))
D$sex = factor(as.character(a$sex))
```

# Exploratory Data Analysis

Devide the study sample into 5 age groups:

<= 40s, 50s, 60s, 70s and >= 80s

## Kaplan-Meier Curves

Then we shows the gender-specific and age-specific Kaplan-Meier curves for the case survival probabilities. The Kaplan–Meier estimator, also known as the product limit estimator, is a non-parametric statistic used to estimate the survival function from lifetime data. The estimator of the survival function $S(t)$ (the probability that life is longer than $t$ is given by:

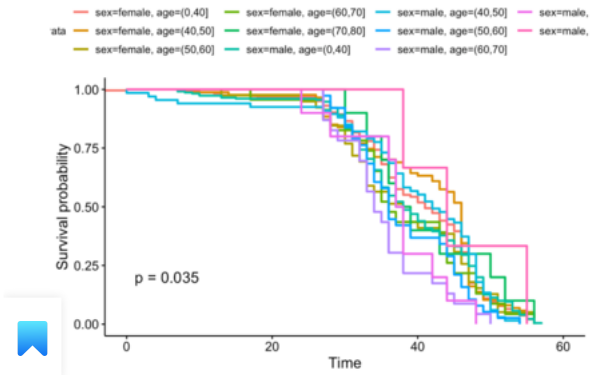
$$\widehat{S}(t) = \prod_{i:\, ti \le t} \left(1 - \frac{di}{ni}\right),$$

with $t_i$ a time when at least one event happened, $d_i$ the "number of events" (e.g., deaths) that happened at time $t_i$, and $n_i$ the "individuals known to have survived" (have not yet had an event or been censored) up to time $t_i$.

```
D = D[complete.cases(D),]

fit = survfit(Surv(date, status) ~ sex + age, data = D)
summary(fit)$table
ggsurvplot(fit, data = D, pval = TRUE)
```

The figure is as follows:




avatar

From the results above in this section, we can seen that COVID-19 is more dangerous for older patients than younger patients, and females are more likely to survival than males to some extent. But the above KM curve is mixed with age and gender factors. In order to further distinguish the influence of the two, in-depth exploration is needed.

# Proposals for Future Analysis

## Log-rank Test

For the effect of gender, we want to use Log-rank Test, which is a hypothesis test to compare the survival distributions of two samples. It is a nonparametric test and appropriate to use when the data are right skewed and censored.

Consider two groups of patients, e.g., treatment vs. control. Let $1, \ldots, J$ be the distinct times of observed events in either group. Let $N_{1,j}$ and $N_{2,j}$ be the number of subjects "at risk" (who have not yet had an event or been censored) at the start of period $j$ in the groups, respectively. Let $O_{1,j}$ and $O_{2,j}$ be the observed number of events in the groups at time $j$. Finally, define $N_j = N_{1,j} + N_{2,j}$ and $O_j = O_{1,j} + O_{2,j}$.

The null hypothesis is that the two groups have identical hazard functions, $H_0 : h_1(t) = h_2(t)$ . Hence, under $H_0$, for each group $i = 1, 2$, $O_{i,j}$ follows a hypergeometric distribution with parameters $N_j$, $N_{i,j}$, $O_j$. This distribution has expected value $E_{i,j} = N_{i,j}\frac{O_j}{N_j}$ and variance $V_{i,j} = E_{i,j}\left(\frac{N_j - O_j}{N_j}\right)\left(\frac{N_j - N_{i,j}}{N_j - 1}\right)$.

For all $j = 1, \ldots, J$, the logrank statistic compares $O_{i,j}$ to its expectation $E_{i,j}$ under $H_0$. It is defined as

$$Z = \frac{\sum_{j=1}^{J}(O_{i,j} - E_{i,j})}{\sqrt{\sum_{j=1}^{J} V_{i,j}}} \xrightarrow{d} \mathcal{N}(0,1) \quad \text{for} \quad i = 1 \text{ or } 2$$

By the Lyapunov CLT, the distribution of $Z$ converges to that of a standard normal distribution as $J$ approaches infinity and therefore can be approximated by the standard normal distribution for a sufficiently large $J$. An improved approximation can be obtained by equating this quantity to Pearson type I or II (beta) distributions with matching first four moments.

## Cox Proportional Hazards Model

The above mentioned methods Kaplan-Meier curves and logrank tests are examples of univariate analysis. They describe the survival according to one factor under investigation, but ignore the impact of any others.

Additionally, Kaplan-Meier curves and logrank tests are useful only when the predictor variable is categorical (e.g.: treatment A vs treatment B; males vs females). They don't work easily for quantitative predictors such as gene expression, weight, or age.

An alternative method is the Cox proportional hazards regression analysis, which works for both quantitative predictor variables and for categorical variables. Furthermore, the Cox regression model extends survival analysis methods to assess simultaneously the effect of several risk factors on survival time.

The Cox proportional-hazards model (Cox, 1972) is essentially a regression model commonly used statistical in medical research for investigating the association between the survival time of patients and one or more predictor variables. Here we use it to estimate the effects of age, gender and even confirmed data.

## :her Potential Tests

Finally, I will try to conduct Wald test, likelihood-ratio test and the Lagrange multiplier test (also known as the score test). For example, a Wald test on the effect of age (chi-square with 4 degrees of freedom) or on the effect of gender (chi-square with 1 degrees of freedom). I will also try to use random forest to explore which factor (confirmed data, gender and age) is most important for CFD.

---



**本页面最后编辑于2020年5月23日 (星期六) 18:35。**