# 344 Project - Titanic Passengers

Florence Wang (Leader) 3933620
Part I Writing, Data Selection

Shaxuan Luo 69950640
Part I Writing, Data Selection

Yang Lei 39123435
Part I Coding (Stratified Sampling), Data Selection

Yitong Zhang 22072904
Part II, Part I Coding(SRS)&
Background/Data summarises Writing, Data Selection

# Part I:

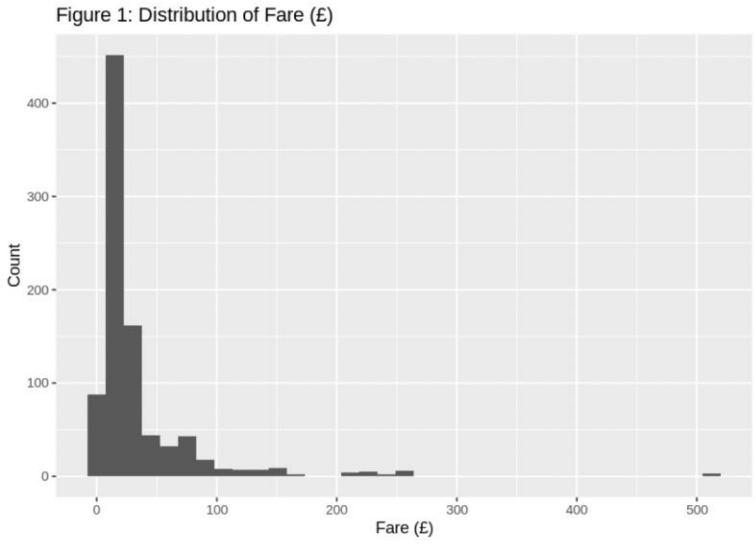
### Background & Objective

People have increased their demand for travel and Cruise tourism today is still seen as a luxurious form of traveling, so what about a hundred years before? The movie Titanic exposed the social class from the rich and famous to the poor and obscure via different levels of tickets. Throughout history, the Titanic is one of the most well-known tragedies, it was one of the largest and most luxurious ships that sank on April 14-15, 1912, killing about more than a thought people. In this report, we'll be specifically looking at the fare on the titanic to understand the fare cost and difference in the early 1900s.

This study aimed to estimate the mean of the fare (population mean) and the proportion of the fare over £35 (population proportion) of 1316 passengers by a data set that was posted by Duong, P. H. (2015). For the sake of our analysis and justification, we will assume 1316 passengers is our population, but this might be biased as discussed in the limitation section.

### Data collection and Data summaries

After filtering out the data for missing values, the data set includes 891 records of the passengers' ID (*PassengerId*), passengers' name (*Name*), passengers' class (*Pclass*: 1 for first class; 2 for second class; 3 for third class), and passengers' fare (*Fare*: in £). The histogram shows the distribution of fare.



The continuous input variable in this dataset is the mean fare of passengers on titanic (Fare), and we will convert it into a binary variable which is the proportion of the fare over £35 to calculate the sample size.

### Simple random sample (SRS)

First, we need to determine the sample size, we demand an accuracy of 5% (m.o.e), 19 times out of 20. We set the cutoff fare rate to be £35 and converted the continuous variable fare to binary variable ***fare_higher_than_mean***. In that way, we found out the proportion P guess is 0.31, and the guessed population variance s-squared guess is 0.2139.

Therefore, we can calculate the sample size with FPC using the formula:

$$n0 = (Z^2_{\alpha/2} * s^2_{guess})/\delta^2 = (1.96^2 * 0.2139)/0.05^2 = 329$$
$$n = n0/(1 + n0/N) = 329/(1 + 329/1316) = 264$$

Hence, we use the function "rep_sample_n()" in R to randomly select 264 individual passengers from the population dataset. Then we calculate the parameter of interest, population proportion, and its 95% confidence interval. (The R code and results are in the Appendix)

### Stratified sampling

We still choose the total sample size of 264 based on the previous sample size calculation, but divide them into three strata (class 1, class 2, and class 3). First, we filter the data with the passengers' class to build three sub-population (class 1, class 2, and class 3). The reason why we split into these three strata is because the passenger's class is clearly associated with the fare rate.

Secondly, we use "rep_sample_n()" to sample each class with the size calculated by optimal allocation,

$$n_h \propto N_h \times \frac{s_{h,guess}}{\sqrt{c_h}}, \qquad h = 1, 2, \cdots, H.$$

($n_h$ is sub-sample size, $N_h$ is sub-population size, $s_{h,guess}$ is the standard deviation of each stratum, $c_h$ is the cost of the sampling). We assume $c_h$ is equal across each stratum, and usually, $n_h = N_h * k * s_{h,guess}/\text{sqrt}(c_h)$, k is some constant.

Therefore, we can calculate the proportion of each stratum and then times 264 (total sample size) to get each stratum sample size. In another word, we need to sample more if the stratum size is large, and the data has larger variance within-strum. Finally, combine three sub-sample to build the sample to estimate the population proportion and its 95% confidence interval.

The reason that choosing optimal allocation as the method to calculate sample size is due to the not equal variance. The first class has a variance of 78.38^2=6143.4828, the second class is 13.417^2=180.016, and the third is 11.77^2=138.7246.

**Results**

Simple Random Sample:

- The estimated mean fare is 34.3164
- The standard error of the estimated mean fare is 2.6902
- The 95% confidence interval for the mean fare is (29.0436, 39.5892)
- The estimated proportion of the fare over £35 is 0.25 (25%)
- The standard error of the estimated proportion of the fare over £35 is 0.0239
- The 95% confidence interval for the proportion of the fare over £35 is (0.2246, 0.3954)

Stratified Sampling

- the estimated mean fare is 34.502
- the standard error of the estimated mean fare is 1.3164
- the 95% confidence interval for the mean fare is (31.922, 37.0823)
- The estimated proportion of the fare over £35 is 0.2607 (26.07%)
- The standard error of the estimated proportion of the fare over £35 is 0.0246
- The 95% confidence interval for the proportion of the fare over £35 is (0.212, 0.309)

By interpreting, for both simple random sampling and stratified sampling, the confidence interval computed in this way will include the true value of the mean fare/the proportion of the fare over £35 for 95% repeated samples.

**Data Analysis & Discussion**

For the advantage and disadvantages of Simple Random Sampling (SRS) and stratified sampling. Usually, the SRS sampling method is easier to compute than the stratified sampling. While Stratified Sampling provides a more equal chance for each possible when constituting the sample from the entire population than SRS. In the calculation process, we include the FPC because our sample size is relatively small, and also we assumed that the population size is known.

In our study, as mentioned in the result, the sample constitution of Stratified Sampling has a smaller variance than the data from SRS. The standard error of mean by SRS is 2.69; The standard error of proportion by SRS is 0.0239; The standard error of mean by Stratified is 1.3164; The standard error of proportion by Stratified is 0.0246.

The smaller standard error of the estimated proportion gives a more accurate estimation. Same to standard error, the confidence interval also provides significant evidence. For the mean, the stratified provides a significantly smaller variance than SRS. So, for the mean, Stratified is better and provides more accurate results. As for the proportion of fare over £35, the results are similar. Thus, for the Stratified, Simple Random Sampling is better than Stratified Sampling since SRS is easier but did not provide a significantly inaccurate result.

**Conclusion & Limitations**

This study aims at the mean of the fare and the proportion of the fare over £35 in 891 records among a total of 1316 passengers. To estimate the mean and percentage over £35 of the fare, two methods are used: Simple Random Sampling (SRS) and Stratified Sampling. The intuition is the stratified sample would be better under our context of clearly seeing three different group, so that we will have a larger between-strata variance and a smaller within-strata.

The results of two different sampling methods are not very different, as previously mentioned in the discussion, SRS is a better method if the variance of Stratified is not significantly smaller than SRS. In conclusion, for the mean, Stratified is better, while for the proportion, SRS is better.

At the same time, there are still some limitations in this study that need to mention. The data that was used as the population did not include all passengers on the Titanic. Therefore, the result may change if there are more passengers recorded. Also, we've noticed that some extreme values in the dataset could make the sample means and variance not representative.

**PART II:**

In the past years, some statisticians criticized the likelihood ratio tests (LRT) as having several testing problems in the multiparameter hypothesis testing problems. However, the inappropriate inference problems were caused by the superior test rather than the LRT. Moreover, this brought up some often ignored problems that some powerful size tests may be scientifically inappropriate and the standard of unbiasedness and admissibility should also be reconsidered. Lastly, the paper disagrees with ignoring "intuition", and believes this could damage the reputation of statistical science.

**Reference:**

Duong, P. H. (2015). https://github.com/datasciencedojo/datasets/blob/master/titanic.csv

**Appendix**

SRS:

In [31]:
```
library(tidyverse)
library(infer)
library(broom)
```

```
Warning message:
“package ‘broom’ was built under R version 4.1.3”
```

In [32]:
```
pop <- read.csv("titanic.csv")
head(pop, 3)
```

A data.frame: 3 × 12

| | PassengerId | Survived | Pclass | Name | Sex | Age | SibSp | Parch | Ticket | Fare | Cabin | Embarked |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | <int> | <int> | <int> | <chr> | <chr> | <dbl> | <int> | <int> | <chr> | <dbl> | <chr> | <chr> |
| **1** | 1 | 0 | 3 | Braund, Mr. Owen Harris | male | 22 | 1 | 0 | A/5 21171 | 7.2500 | | S |
| **2** | 2 | 1 | 1 | Cumings, Mrs. John Bradley (Florence Briggs Thayer) | female | 38 | 1 | 0 | PC 17599 | 71.2833 | C85 | C |
| **3** | 3 | 1 | 3 | Heikkinen, Miss. Laina | female | 26 | 0 | 0 | STON/O2. 3101282 | 7.9250 | | S |

In [62]:
```
pop_tidy <- pop %>% select(PassengerId, Pclass, Fare) %>% mutate(fare_higher_than_mean = ifelse(Fare > 35, 1, 0))
head(sample)
```

A grouped_df: 6 × 4

| replicate | PassengerId | Pclass | Fare |
|---|---|---|---|
| <int> | <int> | <int> | <dbl> |
| 1 | 836 | 1 | 83.1583 |
| 1 | 679 | 3 | 46.9000 |
| 1 | 129 | 3 | 22.3583 |
| 1 | 509 | 3 | 22.5250 |
| 1 | 471 | 3 | 7.2500 |
| 1 | 299 | 1 | 30.5000 |

In [64]:
```
set.seed(1)
sample <- rep_sample_n(pop_tidy, 264, replace = FALSE)
```

In [65]:
```
head(sample)
```

A grouped_df: 6 × 5

| replicate | PassengerId | Pclass | Fare | fare_higher_than_mean |
|---|---|---|---|---|
| <int> | <int> | <int> | <dbl> | <dbl> |
| 1 | 836 | 1 | 83.1583 | 1 |
| 1 | 679 | 3 | 46.9000 | 1 |
| 1 | 129 | 3 | 22.3583 | 0 |
| 1 | 509 | 3 | 22.5250 | 0 |
| 1 | 471 | 3 | 7.2500 | 0 |
| 1 | 299 | 1 | 30.5000 | 0 |

In [54]:
```
# estimate mean of fare
mean_fare <- mean(sample$Fare)
mean_fare
```

34.3163825757576

In [66]:
```
# estimate standard errors of fare
n <- 264
N <- nrow(pop)

se_mean_fare <- sqrt((1-n/N)*var(sample$Fare)/n)
se_mean_fare
```

2.69019502403047

In [67]:
```
sample_summarized <- sample %>% summarize(mean_fare , se_mean_fare)
sample_summarized
```

A tibble: 1 × 3

| replicate | mean_fare | se_mean_fare |
|---|---|---|
| <int> | <dbl> | <dbl> |
| 1 | 34.31638 | 2.690195 |

In [68]:
```
# Calculate the confidence interval
ci_mean_fare <- c(lower = mean_fare - 1.96*se_mean_fare,
                  upper =  mean_fare + 1.96*se_mean_fare)
ci_mean_fare
```

**lower:** 29.0436003286578 **upper:** 39.5891648228573

In [69]:
```
# Summarize the mean, standard error, and variance for binary variable fare_higher_than_mean
sample_summarized_prop <- sample %>% summarize(higher_rate_proportion =mean(sample$fare_higher_than_mean),
                                               se_higher_rate_proportion = sqrt((1-n/N)*fare_higher_rate*(1-fare_higher_rate)/n),
                                              var_higher_rate_proportion=se_higher_rate_proportion^2)
sample_summarized_prop
```

A tibble: 1 × 4

| replicate | higher_rate_proportion | se_higher_rate_proportion | var_higher_rate_proportion |
|---|---|---|---|
| <int> | <dbl> | <dbl> | <dbl> |
| 1 | 0.25 | 0.02387802 | 0.0005701599 |

In [49]:
```
# Confidence interval for binary variable fare_higher_than_mean
ci_fare_higher_rate <- c(lower = fare_higher_rate - 1.96*se_fare_higher_rate,
                         upper = fare_higher_rate + 1.96*se_fare_higher_rate)
ci_fare_higher_rate
```

**lower:** 0.224589580808912 **upper:** 0.395410419191088

# Stratified Sampling:

In [2]:
```
library(tidyverse)
library(infer)
library(dplyr)
```

In [3]:
```
titanic <- read.csv('titanic.csv')
```

In [4]:
```
titanic <- titanic %>% select(PassengerId, Pclass, Fare)
```

In [5]:
```
head(titanic)
nrow(titanic)
```

A data.frame: 6 × 3

| | PassengerId | Pclass | Fare |
|---|---|---|---|
| | <int> | <int> | <dbl> |
| **1** | 1 | 3 | 7.2500 |
| **2** | 2 | 1 | 71.2833 |
| **3** | 3 | 3 | 7.9250 |
| **4** | 4 | 1 | 53.1000 |
| **5** | 5 | 3 | 8.0500 |
| **6** | 6 | 3 | 8.4583 |

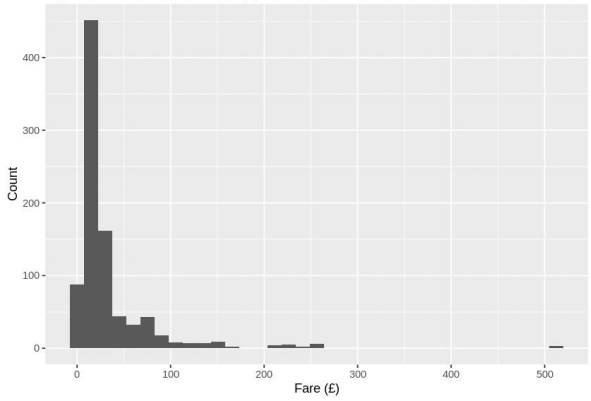
891

In [6]:
```
options(repr.plot.width = 7, repr.plot.height = 5)

fare_dist <- titanic %>% ggplot() +
    geom_histogram(aes(x = Fare), bins = 35) +
    ggtitle('Figure 1: Distribution of Fare (£)') +
    labs(x = 'Fare (£)', y = 'Count')

fare_dist
```

Figure 1: Distribution of Fare (£)



In [7]:
```
N_1 <- titanic %>% filter(Pclass == 1) %>% nrow()
N_2 <- titanic %>% filter(Pclass == 2) %>% nrow()
N_3 <- titanic %>% filter(Pclass == 3) %>% nrow()

prop_1 <- N_1 / nrow(titanic)
prop_2 <- N_2 / nrow(titanic)
prop_3 <- N_3 / nrow(titanic)

N_1
prop_1
N_2
prop_2
N_3
prop_3
```

216

0.242424242424242

184

0.206509539842873

491

0.551066217732884

In [8]:
```
class1 <- titanic %>% filter(Pclass == 1)

var_fare_class1 <- var(class1$Fare)

sd_fare_class1 <- sd(class1$Fare)
sd_fare_class1

optimal_ratio_1_unstandardized <- N_1*sd_fare_class1
optimal_ratio_1_unstandardized
```

78.3803726467288

16930.1604916934

In [9]:
```
class2 <- titanic %>% filter(Pclass == 2)

var_fare_class2 <- var(class2$Fare)

sd_fare_class2 <- sd(class2$Fare)
sd_fare_class2

optimal_ratio_2_unstandardized <- N_2*sd_fare_class2
optimal_ratio_2_unstandardized
```

13.4173987561493

2468.80137113148

In [10]:
```
class3 <- titanic %>% filter(Pclass == 3)

var_fare_class3 <- var(class3$Fare)

sd_fare_class3 <- sd(class3$Fare)
sd_fare_class3

optimal_ratio_3_unstandardized <- N_3*sd_fare_class3
optimal_ratio_3_unstandardized
```

11.7781417043873

5783.06757685417

In [11]:
```
x <- 1/(optimal_ratio_1_unstandardized +
        optimal_ratio_2_unstandardized +
        optimal_ratio_3_unstandardized)

optimal_ratio_1 <- x*optimal_ratio_1_unstandardized
optimal_ratio_2 <- x*optimal_ratio_2_unstandardized
optimal_ratio_3 <- x*optimal_ratio_3_unstandardized

optimal_ratio_1
optimal_ratio_2
optimal_ratio_3
```

0.672311202409157

0.0980382211467601

0.229650576444083

In [12]:
```
sample_size <- 264
optimal_size_1 <- (sample_size * optimal_ratio_1) %>% round()
optimal_size_2 <- (sample_size * optimal_ratio_2) %>% round()
optimal_size_3 <- (sample_size * optimal_ratio_3) %>% round()

optimal_size_1
optimal_size_2
optimal_size_3
```

177

26

61

In [13]:
```
set.seed(1)
sample_c1 <- class1 %>%
                rep_sample_n(size = optimal_size_1, replace = FALSE) %>%
                ungroup() %>%
                select(-replicate)
sample_c2 <- class2 %>% rep_sample_n(size = optimal_size_2, replace = FALSE) %>%
                ungroup() %>%
                select(-replicate)
sample_c3 <- class3 %>% rep_sample_n(size = optimal_size_3, replace = FALSE) %>%
                ungroup() %>%
                select(-replicate)

sample_str <- rbind(sample_c1, sample_c2, sample_c3)
head(sample_str)
```

A tibble: 6 × 3

| PassengerId | Pclass | Fare |
|---|---|---|
| <int> | <int> | <dbl> |
| 310 | 1 | 56.9292 |
| 691 | 1 | 57.0000 |
| 541 | 1 | 71.0000 |
| 670 | 1 | 52.0000 |
| 225 | 1 | 90.0000 |
| 62 | 1 | 80.0000 |

In [14]:
```
N <- nrow(titanic)
N_h <- tapply(titanic$Fare, titanic$Pclass, length)
n_h <- tapply(sample_str$Fare, sample_str$Pclass, length)

fare_h <- tapply(sample_str$Fare, sample_str$Pclass, mean)
var_fare_h <-  tapply(sample_str$Fare, sample_str$Pclass, var)
se_fare_h  <-  sqrt((1 - n_h / N_h) * var_fare_h / n_h)

rbind(fare_h, se_fare_h)

fare_str <- sum(N_h / N * fare_h)
se_str <- sqrt(sum((N_h / N)^2 * se_fare_h^2))
c(fare_str, se_str)
```

A matrix: 2 × 3 of type dbl

| | 1 | 2 | 3 |
|---|---|---|---|
| fare_h | 83.358758 | 20.008173 | 18.440777 |
| se_fare_h | 2.556112 | 1.590579 | 2.021567 |

34.5021515429394 · 1.31640034875389

In [15]:
```
ci_mean_fare_str <- c(lower = fare_str - 1.96*se_str,
                      upper = fare_str + 1.96*se_str)
ci_mean_fare_str
```

**lower:** 31.9220068593818 **upper:** 37.082296226497

In [16]:
```
sample_str <- sample_str %>%
                mutate(fare_higher_than_mean = ifelse(Fare > 35, 1, 0))

head(sample_str, 3)
slice(sample_str, 179:181)
tail(sample_str, 3)
```

A tibble: 3 × 4

| PassengerId | Pclass | Fare | fare_higher_than_mean |
|---|---|---|---|
| <int> | <int> | <dbl> | <dbl> |
| 310 | 1 | 56.9292 | 1 |
| 691 | 1 | 57.0000 | 1 |
| 541 | 1 | 71.0000 | 1 |

A tibble: 3 × 4

| PassengerId | Pclass | Fare | fare_higher_than_mean |
|---|---|---|---|
| <int> | <int> | <dbl> | <dbl> |
| 99 | 2 | 23.0000 | 0 |
| 609 | 2 | 41.5792 | 1 |
| 638 | 2 | 26.2500 | 0 |

A tibble: 3 × 4

| PassengerId | Pclass | Fare | fare_higher_than_mean |
|---|---|---|---|
| <int> | <int> | <dbl> | <dbl> |
| 385 | 3 | 7.8958 | 0 |
| 501 | 3 | 8.6625 | 0 |
| 525 | 3 | 7.2292 | 0 |

In [17]:
```
higher_fare_rate_h <- tapply(sample_str$fare_higher_than_mean, sample_str$Pclass, mean)
var_higher_fare_h <- higher_fare_rate_h*(1-higher_fare_rate_h)
se_higher_fare_h  <-  sqrt((1 - n_h / N_h) * var_higher_fare_h / n_h)

rbind(higher_fare_rate_h, se_higher_fare_h)

higher_fare_rate_str <- sum(N_h / N * higher_fare_rate_h)
se_higher_fare_rate_str <- sqrt(sum((N_h / N)^2 * se_higher_fare_h^2))
c(higher_fare_rate_str, se_higher_fare_rate_str)
```

A matrix: 2 × 3 of type dbl

| | 1 | 2 | 3 |
|---|---|---|---|
| higher_fare_rate_h | 0.71186441 | 0.07692308 | 0.13114754 |
| se_higher_fare_h | 0.01446492 | 0.04842618 | 0.04044657 |

0.260729518115829 · 0.0246798237922776

In [18]:
```
ci_fare_higher_rate <- c(lower = higher_fare_rate_str - 1.96*se_higher_fare_rate_str,
                         upper = higher_fare_rate_str + 1.96*se_higher_fare_rate_str)
ci_fare_higher_rate
```

**lower:** 0.212357063482965 **upper:** 0.309101972748693