Final Project

Hafizh Rahmatdianto Yusuf

2023-12-01

Load data and packages

```
options(warn = -1)

# Load packages
set.seed(5)
library(moderndive)
library(skimr)
library(ggplot2)
suppressPackageStartupMessages(library(dplyr))
library(infer)
library(readxl)

# Load data
suppressWarnings({
  hospital_data <- read_excel("~/Documents/UMich/Health Informatics/Fall 2023/SI544/Final_Project/hospi
     col_types = c("date", "numeric", "numeric",
         "numeric", "text", "text", "text",
        "date", "date", "date", "text"))
})
```

Here we use dataset from Kaggle called `Hospital patient data` which can be accessed here https://www.kaggle.com/datasets/abdulqaderasiirii/hospital-patient-data/. This dataset contains hospital patient data that covers the expenses, revenue generated, and the financial class of the patients. The dataset has granularity of `Date` and `Patient ID`.

Data exploration

```
# Check data samples
glimpse(hospital_data)
```

```
## Rows: 29,998
## Columns: 11
## $ Date                    <dttm> 2019-11-04, 2019-11-06, 2019-11-02, 2019-11-~
## $ `Medication Revenue`    <dbl> 1183.22, 738.48, 660.00, 600.00, 591.60, 586.~
## $ `Lab Cost`              <dbl> 10.0, NA, NA, NA, NA, NA, 92.5, NA, NA, 10.0,~
```

```
## $ `Consultation Revenue`   <dbl> 20.17, 15.00, 21.17, NA, 12.00, 13.00, 15.00,~
## $ `Doctor Type`            <chr> "ANCHOR", "ANCHOR", "ANCHOR", "ANCHOR", "ANCH~
## $ `Financial Class`        <chr> "HMO", "INSURANCE", "HMO", "MEDICARE", "INSUR~
## $ `Patient Type`           <chr> "OUTPATIENT", "OUTPATIENT", "OUTPATIENT", "OU~
## $ `Entry Time`             <dttm> 1899-12-31 08:35:45, 1899-12-31 19:19:16, 18~
## $ `Post-Consultation Time` <dttm> 1899-12-31 09:17:54, 1899-12-31 21:02:36, 18~
## $ `Completion Time`        <dttm> 1899-12-31 09:29:46, 1899-12-31 21:24:07, 18~
## $ `Patient ID`             <chr> "C10001", "C10002", "C10003", "C10004", "C100~
```

```
# Check distinct values of categorical data
distinct_values <- sapply((hospital_data %>% select(`Doctor Type`, `Financial Class`, `Patient Type`)),
distinct_values
```

```
## $`Doctor Type`
## [1] "ANCHOR"   "LOCUM"    "FLOATING"
##
## $`Financial Class`
## [1] "HMO"       "INSURANCE" "MEDICARE"  "CORPORATE" "PRIVATE"
##
## $`Patient Type`
## [1] "OUTPATIENT"
```

Here we focus on exploring data. As seen above we have 11 columns which consists of 3 numerical data, 4 string data, and 4 date data. Among the 4 string data, there are 3 categorical value which is `Doctor Type`, `Financial Class`, and `Patient Type`. After doing further research for each categorical value, here is the definition of each categorical value:

`Doctor Type`

- Anchor = Permanent physicians
- Locum = Temporary or substitute physician
- Floating = Physician without permanent location

`Financial Class`

- HMO = Health Maintenance Organization, a type of health insurance plan that typically requires members to choose a primary care physician
- Insurance = Healthcare insurance coverage
- Medicare = Federal health insurance program
- Corporate = Healthcare insurance plans provided by corporation for their employees
- Private = Private insurance plans

As for the data range, it spans of 13 days worth of data from 1-13 November 2019. Another interesting data points are `Medication Revenue` and `Lab Cost`. The `Medication Revenue` is defined as the revenue got from medication, while `Lab Cost` is the cost paid by the patient for laboratory services.

Main research question

Using this dataset, we formulate two main research questions:

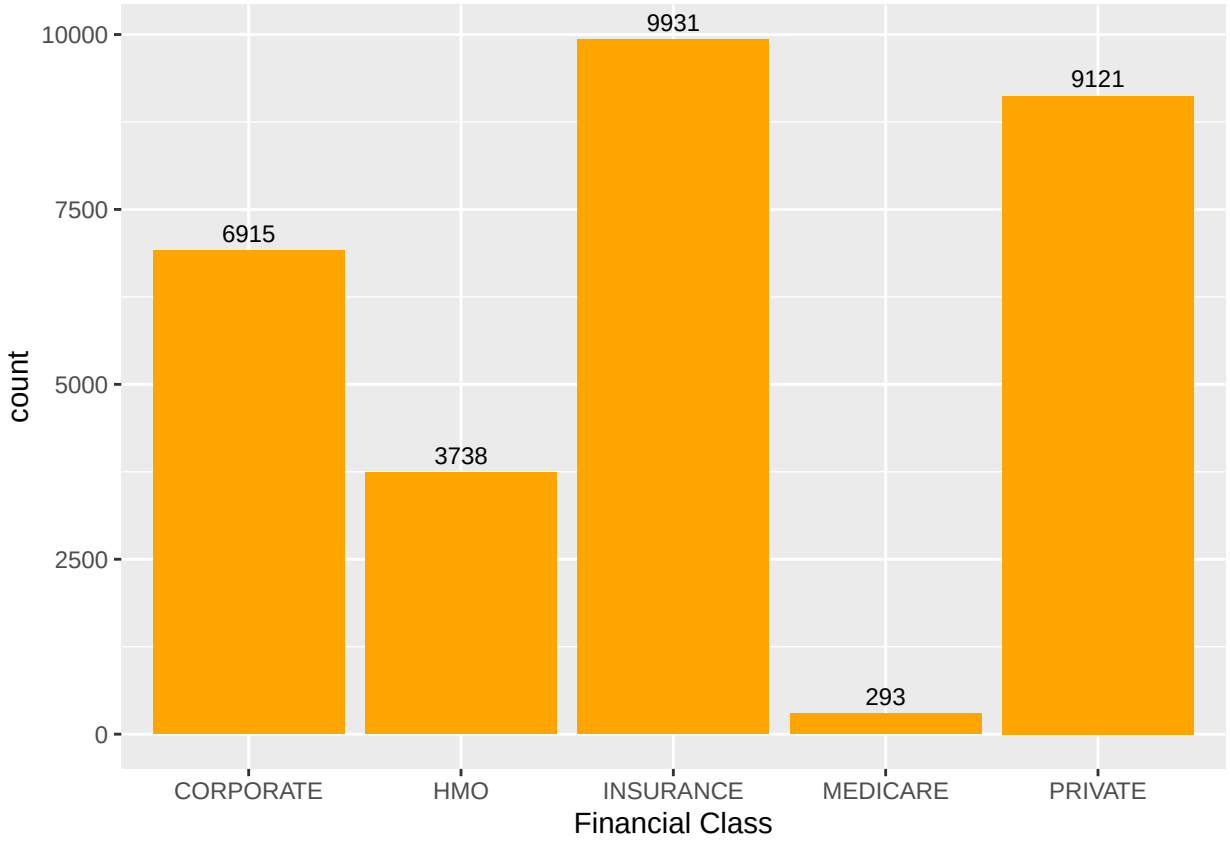
1. What is the relationship between `Medication Revenue` and `Lab Cost` and how `Financial Class` affects it
2. How does `Financial Class` affects the waiting time of the patients.

Since there's multiple `Financial Class`, we tried to narrow the class down into two classes, federal insurance (Medicare) and private insurance (other than Medicare). Those two classes will be called `Insurance Type`. While for the waiting time we calculate a new variable called `Waiting Time` which will be defined as `Post-Consultation Time` - `Entry Time`. In summary, we are going use those two new variables, `Insurance Type` and `Waiting Time`, for answering our research questions.

Data visualization

Financial class data distribution

```
ggplot(hospital_data, aes(x=`Financial Class`)) +
  geom_bar(fill = "orange")+
  geom_text(stat = "count", aes(label = stat(count)),
            vjust = -0.5, color = "black", size = 3)
```

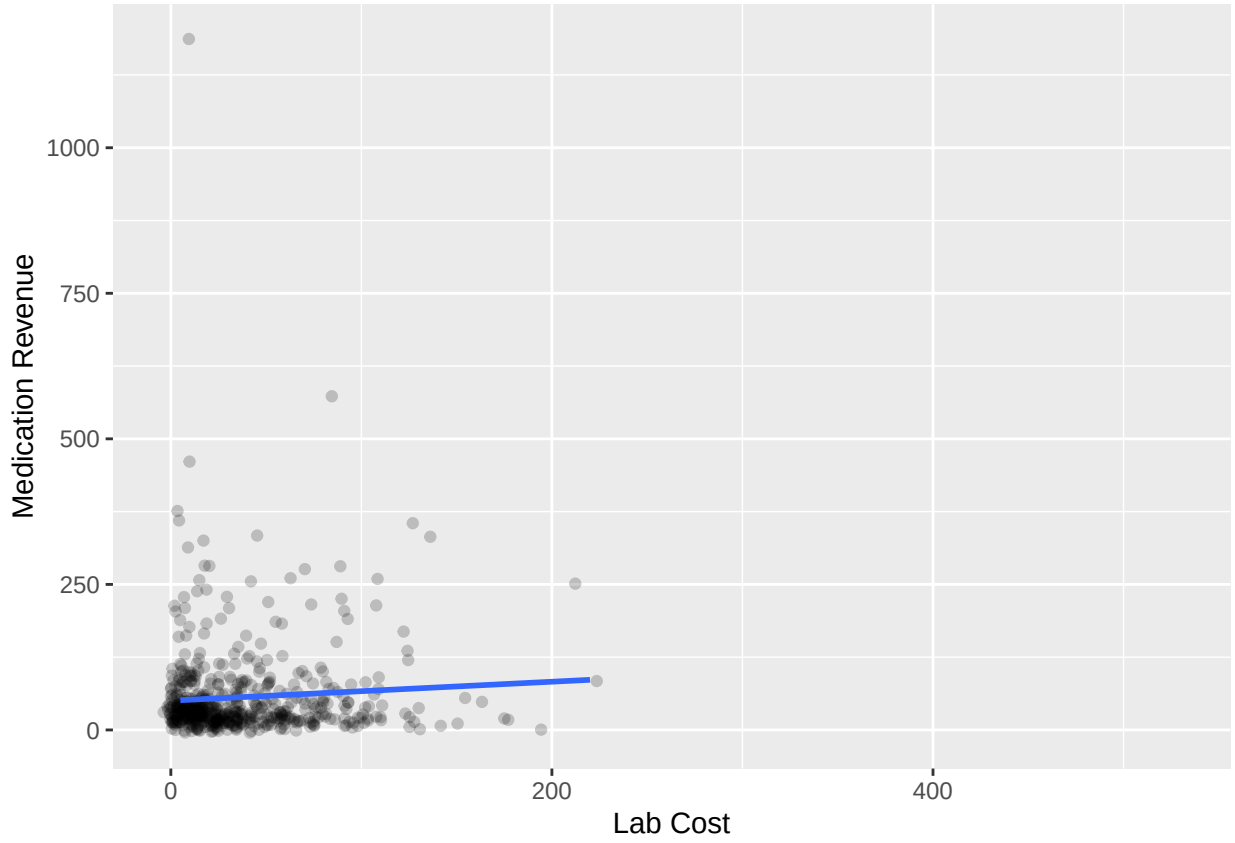


As seen above, the `Financial Class` data distribution, most of the data are from the INSURANCE, while MEDICARE has the least number of data. This distribution visualizes the potential bias of the analysis result later on.

Medication revenue comparison with lab cost

```
ggplot(hospital_data, aes(x=`Lab Cost`, y=`Medication Revenue`)) +
  geom_jitter(alpha=0.2, width = 10, height = 10) +
  geom_smooth(method = "lm", se = FALSE)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



As for the scatterplot above, it describes how the relationship between `Medication Revenue` and `Lab Cost`. On first glance, the linear line shows positive relationship between those two variables, but detailed analysis will be performed below.

Data wrangling

```
# Add necessary columns of new classification
hd_clone <- hospital_data
hd_clone <- hd_clone %>% mutate(`Waiting Time` = `Post-Consultation Time` - `Entry Time`)
```

```
hd_clone <- hd_clone %>% mutate(`Insurance Type` = ifelse(`Financial Class` == "MEDICARE", "Federal Insu
hd_clone <- hd_clone %>% group_by(`Date`, `Insurance Type`) %>% mutate(`Avg Waiting Time` = mean(`Waiti
```

Here we focus on data wrangling. We created three new columns:

1. `Waiting Time` is the waiting time of the patient
2. `Insurance Type` is the type of insurance (Federal or Non-Federal) used by patients
3. `Avg Waiting Time` is the average waiting time for each `Insurance Type` per-day (or `Date`)

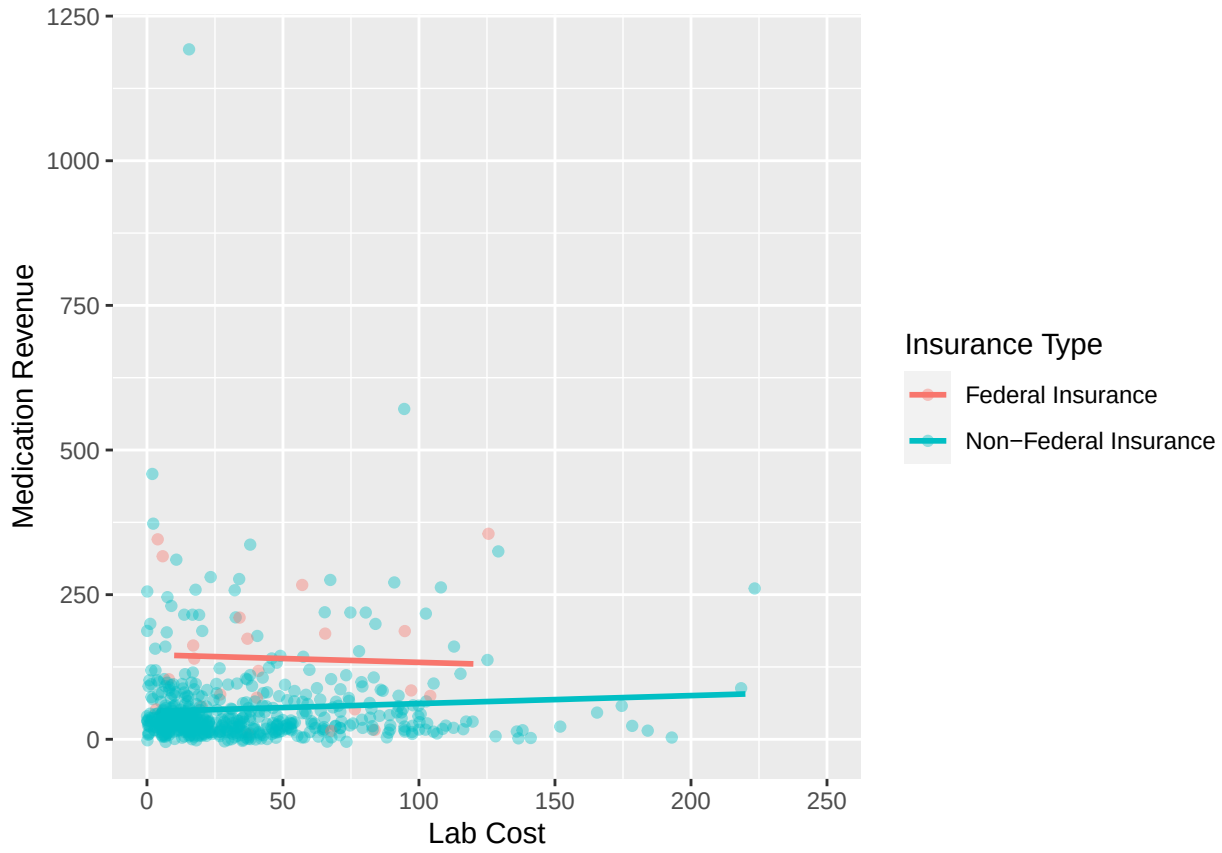
All those new columns will be used for the analysis below.

Multiple regression

Graph

```
ggplot(hd_clone, aes(x=`Lab Cost`, y=`Medication Revenue`, color=`Insurance Type`)) +
  geom_jitter(alpha=0.4, width=10, height=10) +
  geom_smooth(method = "lm", se = FALSE) +
  xlim(0,250)
```

```
## `geom_smooth()` using formula = 'y ~ x'
```



From the graph above we can see that the blue line showing positive relationship, while the red line showing negative relationship. Below we deep dive into the regression table to see how is the linear model actually is.

Linear model

```
model_parallel_slopes <- lm(`Medication Revenue` ~ `Lab Cost` * `Insurance Type`,
                            data = hd_clone)
get_regression_table(model_parallel_slopes)
```

```
## # A tibble: 4 x 7
##   term                     estimate std_error statistic p_value lower_ci upper_ci
##   <chr>                       <dbl>     <dbl>     <dbl>   <dbl>    <dbl>    <dbl>
## 1 intercept                  146.       30.5       4.80    0       86.4    206.
## 2 `Lab Cost`                  -0.133     0.515    -0.258   0.797   -1.14     0.879
## 3 `Insurance Type`Non-Fe~   -98.3       30.8      -3.19    0.002 -159.     -37.8
## 4 `Lab Cost`:`Insurance ~     0.271     0.524     0.517   0.606   -0.759    1.3
```

The equation for the regression lines are as follow:

- y-Federal Insurance = 146.2 - 0.13.`Lab Cost`
- y-Non-Federal Insurance = 47.9 + 0.14.`Lab Cost`

The Federal Insurance has negative slope while the counterpart has positive slope, which aligns with our Initial findings. This means patient with Federal Insurance generate less `Medication Revenue` along with the increase of `Lab Cost` they paid. Meanwhile, patient with Non-Federal Insurance generate more `Medication Revenue` along with the increase of the `Lab Cost` paid.

Hypothesis testing

Hypothesis 1

For the first hypothesis we tried to check whether there is difference betweeen the waiting time of Federal and Non-Federal Insurance. Here are the H0 and HA for this hypothesis testing.

Tf -> Federal Insurance; Tn -> Non-Federal Insurance

H0: Tf = Tn -> Tf - Tn = 0

In words, H0 will be if there is no difference between the waiting time of Federal and Non-Federal Insurance.

HA: Tf != Tn -> Tf - Tn != 0

Alternatively, HA will be if there is difference between waiting time of Federal and Non-Federal Insurance.

For this matter we chose the threshold alpha of 0.05.

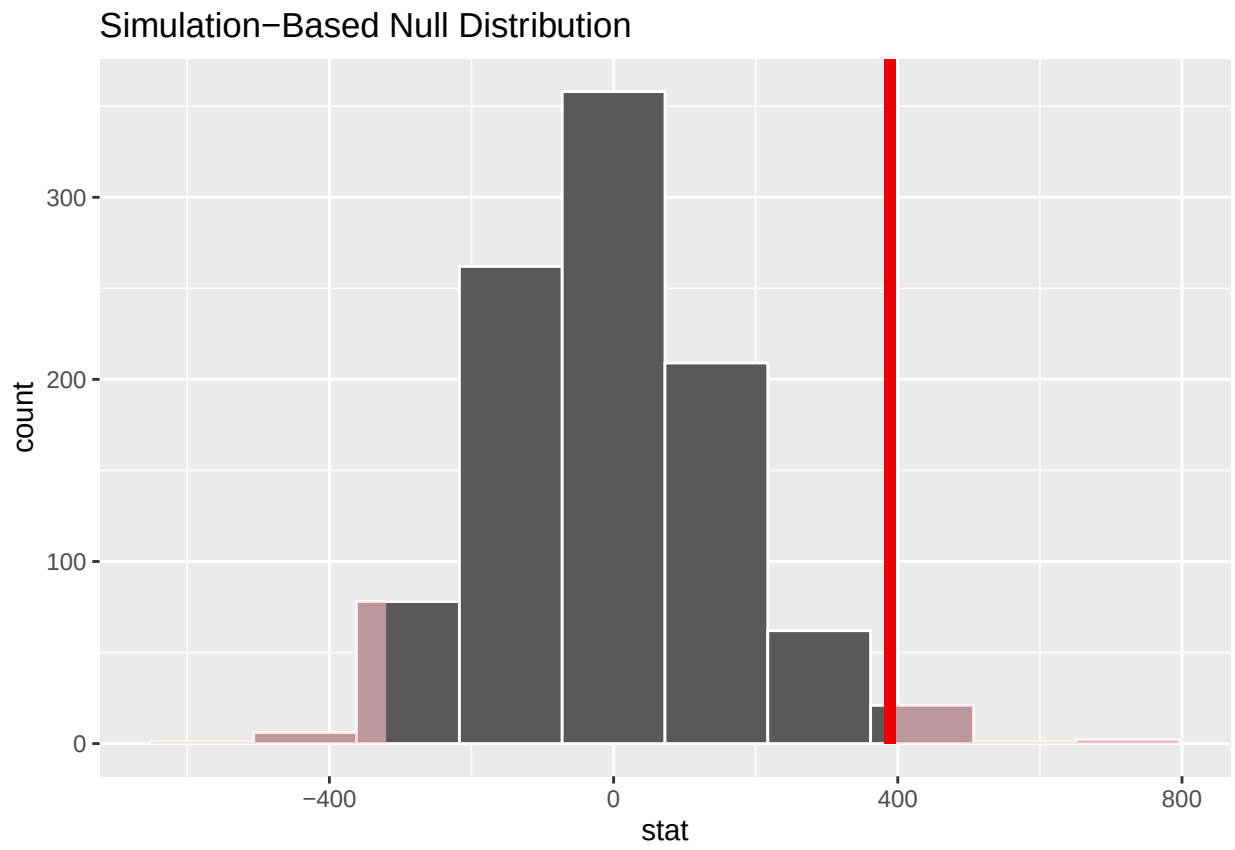
```
# Select data
hd_clone_sel <- hd_clone %>% distinct(`Insurance Type`, `Waiting Time`)
hd_clone_sel$`Waiting Time` <- as.integer(hd_clone_sel$`Waiting Time`)
```

```r
# Define the null distribution of H0
null_distribution_hd <- hd_clone_sel %>%
  specify(formula = `Waiting Time` ~ `Insurance Type`) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("Federal Insurance", "Non-Federal Insurance"))

obs_diff_means <- hd_clone_sel %>%
  specify(formula = `Waiting Time` ~ `Insurance Type`) %>%
  calculate(stat = "diff in means", order = c("Federal Insurance", "Non-Federal Insurance"))
obs_diff_means
```

```
## Response: Waiting Time (numeric)
## Explanatory: Insurance Type (factor)
## # A tibble: 1 x 1
##    stat
##   <dbl>
## 1  389.
```

```r
# Visualize the distribution along with the p-value
visualize(null_distribution_hd, bins = 10) +
  shade_p_value(obs_stat = obs_diff_means, direction = "both")
```



```r
# Get the p-value
null_distribution_hd %>%
  get_p_value(obs_stat = obs_diff_means, direction = "both")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.026
```

Based on the result above, we got p-value of 0.026, which is lower than the threshold alpha we set of 0.05. With this in mind, we reject the H0 and in favor of HA, there is a difference of `Waiting Time` among different `Insurance Type`. The next natural question would be, which one has the higher waiting time? To answer this we also tried it in the second hypothesis testing.

Population mean

```r
hd_clone %>% group_by(`Insurance Type`) %>% summarize(`Avg Waiting Time` = mean(`Waiting Time`, na.rm =
```

```
## # A tibble: 2 x 2
##   `Insurance Type`      `Avg Waiting Time`
##   <chr>                 <drtn>
## 1 Federal Insurance     3018.853 secs
## 2 Non-Federal Insurance 2328.115 secs
```

Since we have the population data, we can try to see the average of the waiting time for each `Insurance Type`. From the stats above we can see that the Federal Insurance has the higher average waiting time compared to Non-Federal Insurance. Then, what if we tried to use the average waiting time for hypothesis testing instead?

Hypothesis 2

For this second hypothesis we tried to check whether there is difference between the average waiting time of Federal and Non-Federal Insurance. Here are the H0 and HA for this hypothesis testing.

ATf -> Federal Insurance; ATn -> Non-Federal Insurance

H0: ATf <= ATn -> ATf - ATn <= 0

Here we tried to change the hypothesis a bit. The previous two had H0 of the waiting time of Federal Insurance is higher or the same compared to the counterpart. Here we twist it into if the waiting time of the Federal Insurance is same or lower than the Non-Federal Insurance, pretty similar to the population estimate we tried earlier.

HA: ATf > ATn -> ATf - ATn > 0

Alternatively, HA will be if waiting time of the Federal Insurance is higher than the Non-Federal Insurance

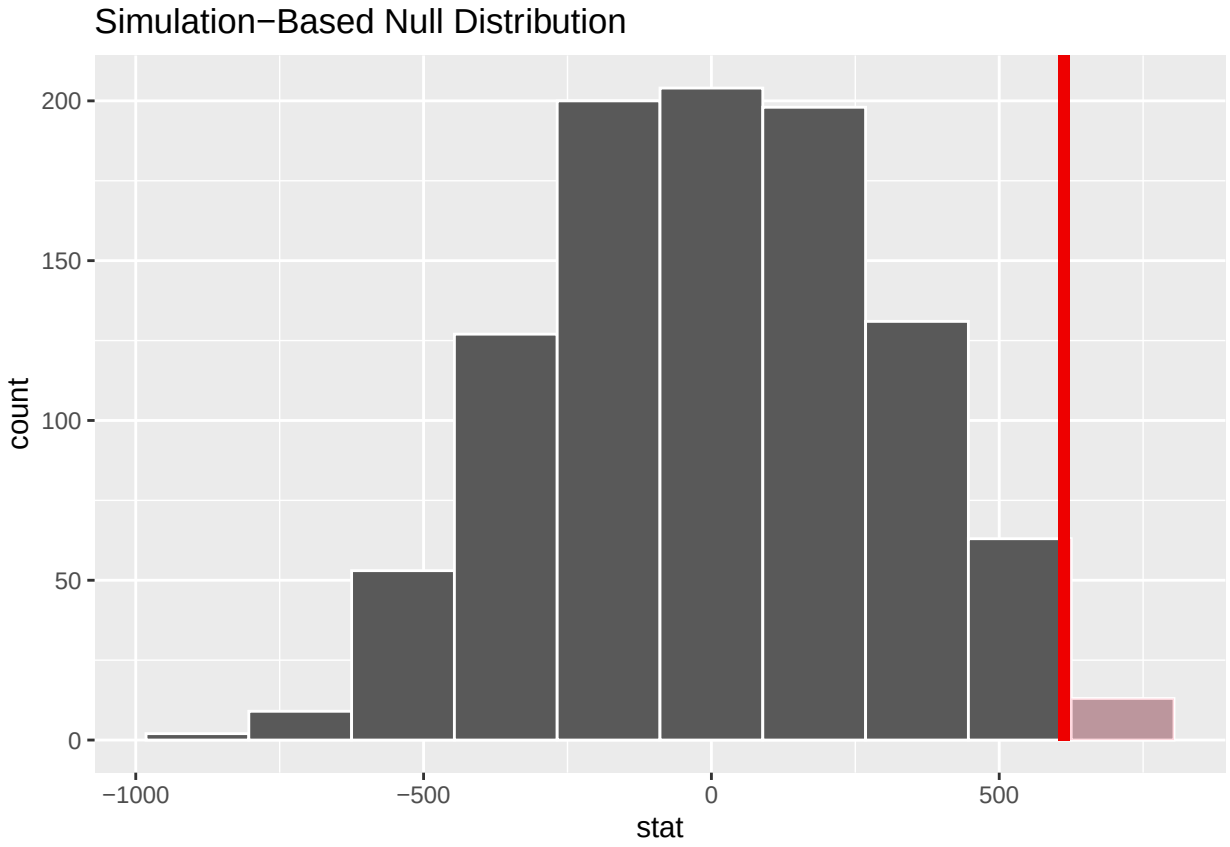
We use threshold alpha of 0.05 again in here.

```r
# Select data
hd_clone_sel_avg <- hd_clone %>% distinct(`Insurance Type`, `Avg Waiting Time`)
hd_clone_sel_avg$`Avg Waiting Time` <- as.integer(hd_clone_sel_avg$`Avg Waiting Time`)
# hd_clone_sel_avg

# Define the null distribution of HO
null_distribution_hd_avg <- hd_clone_sel_avg %>%
  specify(formula = `Avg Waiting Time` ~ `Insurance Type`) %>%
  hypothesize(null = "independence") %>%
  generate(reps = 1000, type = "permute") %>%
  calculate(stat = "diff in means", order = c("Federal Insurance", "Non-Federal Insurance"))
# null_distribution_hd_avg

obs_diff_means_avg <- hd_clone_sel_avg %>%
  specify(formula = `Avg Waiting Time` ~ `Insurance Type`) %>%
  calculate(stat = "diff in means", order = c("Federal Insurance", "Non-Federal Insurance"))

# Visualize the distribution along with the p-value
visualize(null_distribution_hd_avg, bins = 10) +
  shade_p_value(obs_stat = obs_diff_means_avg, direction = "right") # Since the HA uses > operator, we
```




```r
# Get the p-value
null_distribution_hd_avg %>%
  get_p_value(obs_stat = obs_diff_means_avg, direction = "right")
```

```
## # A tibble: 1 x 1
##   p_value
##     <dbl>
## 1   0.014
```

The p-value is in this part is low and less than the alpha that we set previously. This signifies that we reject the null hypothesis H0 of Federal Insurance has lower or same average waiting time as the Non-Federal Insurance, instead we favor the hypothesis of Federal Insurance has higher average waiting time compared to the Non-Federal Insurance. This findings is align well with the population mean we found earlier.

Conclusion

From this project of analyzing `Hospital patient data` sourced from Kaggle, we found two answers for our two research questions.

1. What is the relationship between `Medication Revenue` and `Lab Cost` and how `Financial Class` affects it

The relationship of `Medication Revenue` and `Lab Cost` differ for each `Financial Class`. Federal Insurance has negative relationship of revenue and cost which signifies that the higher the revenue, the lower the cost paid by the patients. This might be influenced by the fact that the Federal Insurance is covered by the Federal Government, however further analysis is required using more valid data. While, the Non-Federal Insurance has positive relationship of revenue and cost, signifying a higher revenue if the cost increased. Both linear equation can be seen below.

- y-`Federal Insurance` = 146.2 - 0.13.`Lab Cost`
- y-`Non-Federal Insurance` = 47.9 + 0.14.`Lab Cost`

2. How does `Financial Class` affects the waiting time of the patients.

These are the key findings based on our three hypothesis testing:

- From the first hypothesis testing, we found that there is a difference in waiting time of Federal and Non-Federal Insurance.
- From the population mean perspective, Federal Insurance has higher average waiting time compared to the Non-Federal Insurance. This might be caused by the data point of Federal Insurance is far less than the counterpart, 293 Federal Insurance data compared to 29705 Non-Federal Insurance data.
- From the second hypothesis testing, we found that the average waiting time of Federal Insurance is higher than the Non-Federal Insurance.

Improvement

This project can be improved with:

1. More data on the Federal Insurance part
2. Clearer definition for each data points within the dataset
3. Extended data span beyond 13 days