First Line Cancer Chemotherapy Analysis with AB Test

Zeyu Li

January 18th, 2025

```
# Install Packages -----------------------------
library("VennDiagram")
```

```
## Loading required package: grid
```

```
## Loading required package: futile.logger
```

```
library("ggplot2")
library("dplyr")
```

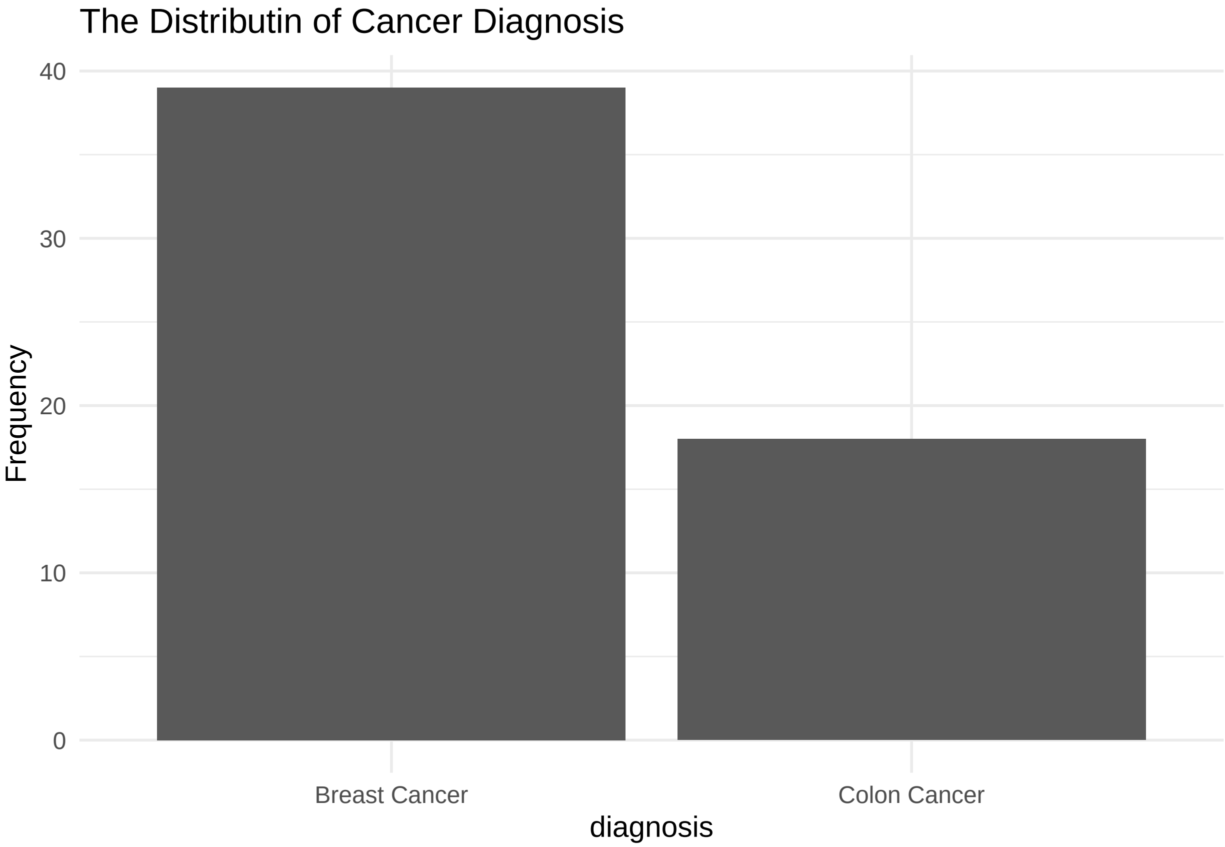
```
##
## Attaching package: 'dplyr'
```

```
## The following objects are masked from 'package:stats':
##
##     filter, lag
```

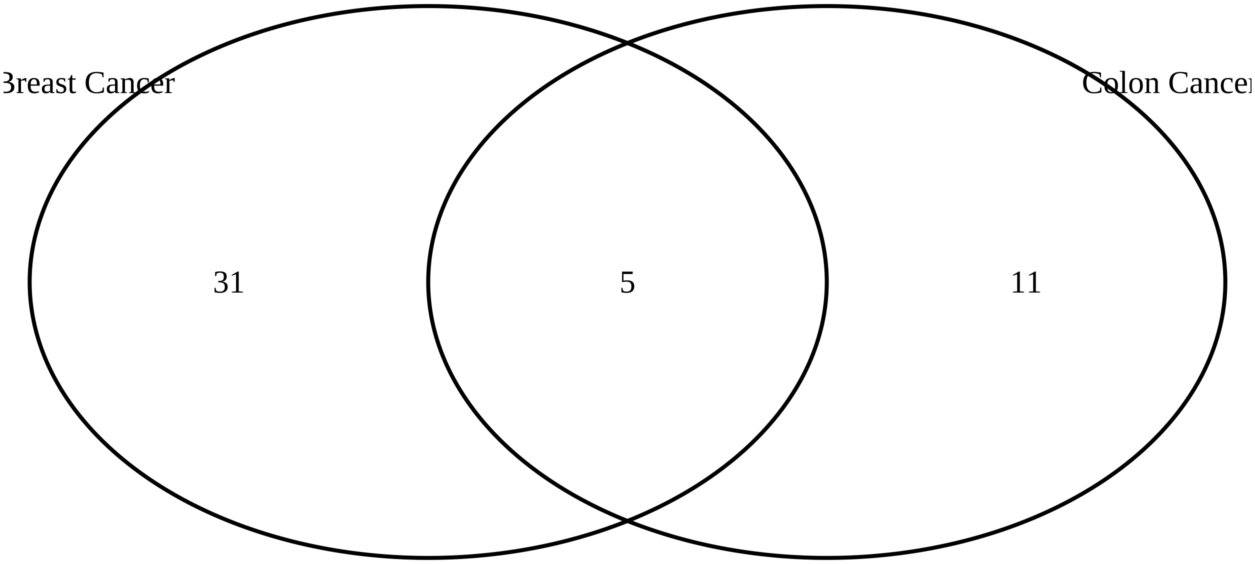
```
## The following objects are masked from 'package:base':
##
##     intersect, setdiff, setequal, union
```

```
library("tidyr")
library("broom")

# Load Data ------------------------------------
diag <- read.csv("Patient_Diagnosis_(2).csv", header=TRUE)
trt <- read.csv("Patient_Treatment_(2).csv", header=TRUE)

##### General Questions #####

# 1. What are the audiences for this analysis project? Do they come from the technical
# background or not?

# 2. In addition to the information provided in the data set, are there any other
# factors that might influence which type of treatment administrated?

# 3. What are the demographic information of the patients? For example, do we have
# the information like age, gender etc... Will we perform the subgroup analysis
# later for this cohort?

##### Data Analysis Questions #####

# Question 1 -----------------------------------

# 1.Select the Column of Diagnosis
```

```
df <- diag %>%
  select(diagnosis)

# 2. Create a barplot with ggplot2 for the distribution of cancer type
bplot<- ggplot(df, aes(x = diagnosis)) +
  geom_bar(stat = "count") +
  labs(title = "The Distributin of Cancer Diagnosis", y = "Frequency") +
  theme_minimal() # Create a barplot with ggplot2 for the distribution of cancer type

print(bplot)
```



```
# 3.Create a Venn Diagram for the distribution of cancer type
vplot <- venn.diagram(
  x = list(
    diag %>% filter(diagnosis=="Breast Cancer") %>% select(patient_id) %>% unlist() ,
    diag %>% filter(diagnosis=="Colon Cancer") %>% select(patient_id) %>% unlist()
  ),
    category.names = c("Breast Cancer" , "Colon Cancer"),
    height = 300 ,
    width = 300 ,
    res = 300,
    filename = NULL,
    scaled = FALSE,
    output = TRUE,cat.default.pos = "outer")

grid.newpage()  # Create a new page
grid.draw(vplot)
```

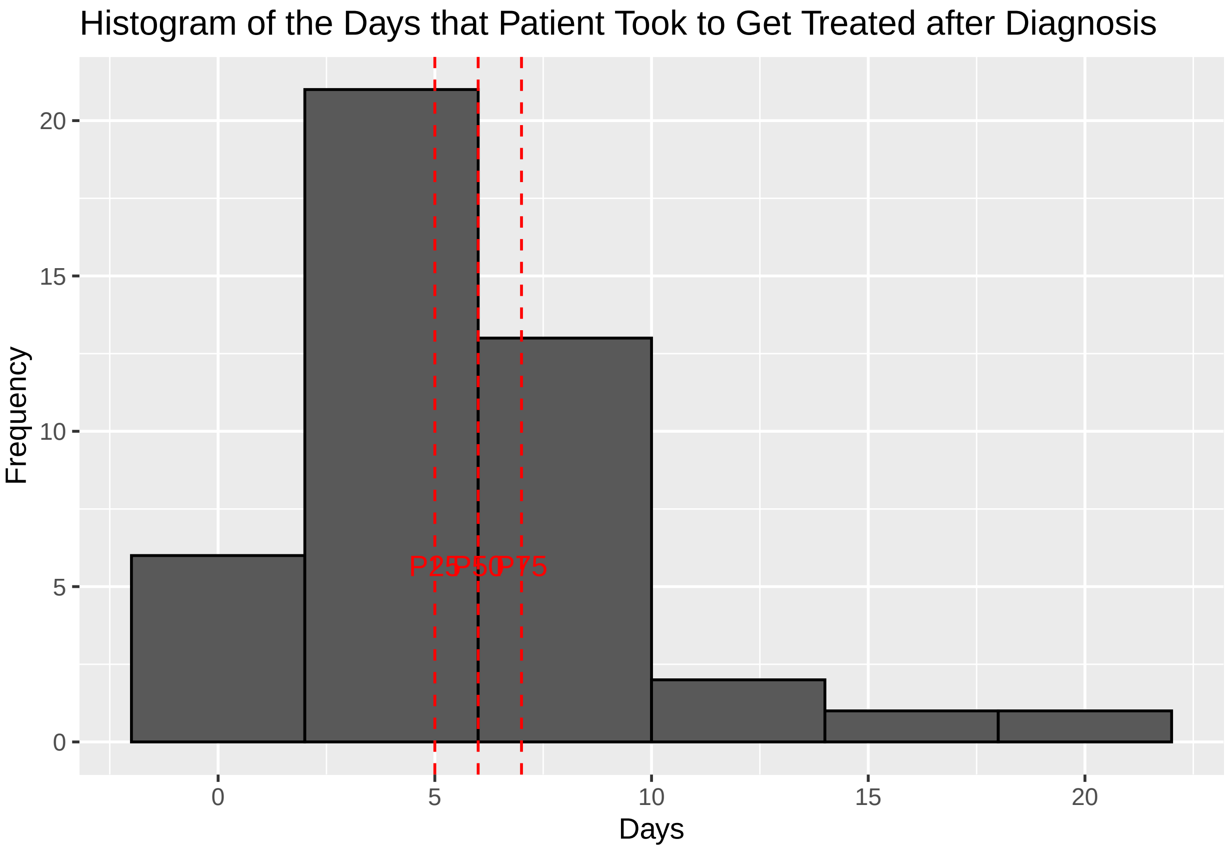


```
# Thought Process: according to the information provided on the question, we know that
# patients can be diagnosed with more than one cancer. It means that it's likely there
# is an overlap between the patients with breast cancers and patients with colon cancer.
# So I not only generate the barplot for the distribution, but also create a Venn diagram
# to show the overlap between the two populations.

# Question 2 -----------------------------------
# 1.Convert the date variables to date format
trt <- trt %>%
  mutate(daten = as.Date(treatment_date, format = "%m/%d/%y"))

# We only kept the initial diagnosis for patient with two types of cancer to determine
# the duration between diagnosis and the start of therapy because we are lack of the
# information regarding the specific treatment administered for each cancer type.
diag_flt <- diag %>%
  mutate(diagnosis_date = as.Date(diagnosis_date, format = "%m/%d/%y")) %>%
  group_by(patient_id) %>%
  summarise(diagnosis_date = min(diagnosis_date))

# 2.Get first date of treatment
min_dates <- trt %>%
  group_by(patient_id) %>%
  summarise(min_date = min(daten))

# 3.Merge diag and min_dates
merged_data <- diag_flt %>%
    left_join(min_dates, by = 'patient_id') %>%
    mutate(datediff = as.numeric(difftime(min_date, diagnosis_date, units = 'days') + 1))

trt_start <- merged_data %>%
    select(all_of(c("patient_id","diagnosis_date","min_date","datediff"))) %>%
    rename(treament_start_date = min_date, days = datediff)

# 4. Generate histogram for the days
# In order to better present the data in the histogram, the outlier, patient who got
# the treatment before diagnosis, and patients with no treatment are excluded from the plot.
```

```
trt_fil <- trt_start[trt_start$days > 0 & trt_start$days < 300 ,] %>%
    na.omit(trt_start)

percentiles <- quantile(trt_fil$days, probs = c(0.25, 0.50, 0.75),na.rm = TRUE)

histogram <- ggplot(trt_fil, aes(x = days)) +
  geom_histogram(binwidth = 4, color = "black") +
  labs(title = "Histogram of the Days that Patient Took to Get Treated after Diagnosis",
       x = "Days", y = "Frequency") +
  geom_vline(xintercept = percentiles, linetype = "dashed", color = "red") +
  geom_text(data = data.frame(x = percentiles, label = paste0("P", c(25, 50, 75))),
            aes(x = x, label = label), y = 5, vjust = -0.5, color = "red")

# Results:
print(trt_start)
```

```
## # A tibble: 47 x 4
##    patient_id diagnosis_date treament_start_date  days
##         <int> <date>         <date>              <dbl>
##  1       2038 2010-01-21     2010-01-24              4
##  2       2120 2010-01-09     2010-01-23             15
##  3       2175 2010-02-17     2010-02-21              5
##  4       2238 2010-01-21     2010-01-21              1
##  5       2407 2010-06-13     2010-06-19              7
##  6       2425 2010-12-15     2010-12-19              5
##  7       2462 2011-01-07     2011-01-11              5
##  8       2475 2010-02-17     2010-02-17              1
##  9       2607 2010-06-13     2010-07-03             21
## 10       2634 2011-02-19     2011-12-20            305
## # i 37 more rows
```

```
print(histogram)
```



```
# Question 3 ------------------------------------

# First line treatment definition: The first treatment given for a disease. It is often
# part of a standard set of treatments, such as surgery followed by chemotherapy and
# radiation. When used by itself, first-line therapy is the one accepted as the best
# treatment. If it doesn't cure the disease or it causes severe side effects, other
# treatment may be added or used instead.

# Derive the first line treatment
# 1.Select the treatment that patient received on the first day
first_line <- trt %>%
    left_join(min_dates, by='patient_id') %>%
    filter(daten == min_date) %>%
    distinct(patient_id, treatment_date, drug_code,.keep_all=TRUE) %>%
    arrange(patient_id, drug_code) %>%
    group_by(patient_id) %>%
    mutate(n = row_number()) %>%
    mutate(n = as.character(n), patient_id = as.character(patient_id))

# 2.Transpose the data to ensure combination therapy considered
tran_first <- pivot_wider(first_line , names_from = n, values_from = drug_code)
merge_first <- merge(diag, tran_first, by='patient_id', all=TRUE) %>%
     mutate(across(everything(), ~ replace(.x, is.na(.x), ""))) %>%
     unite(drug, "1","2", sep="") %>%
     distinct(patient_id, diagnosis_date, diagnosis, daten, drug , .keep_all = TRUE)

# 3.Create frequency table to show which drug regiments indicated to be used as first-line
```

```r
# of treatment for breast cancer and colon cancer
frequency <- merge_first %>%
  group_by(diagnosis, drug) %>%
  summarise(frequency = n()) %>%
  arrange(diagnosis, desc(frequency))
```

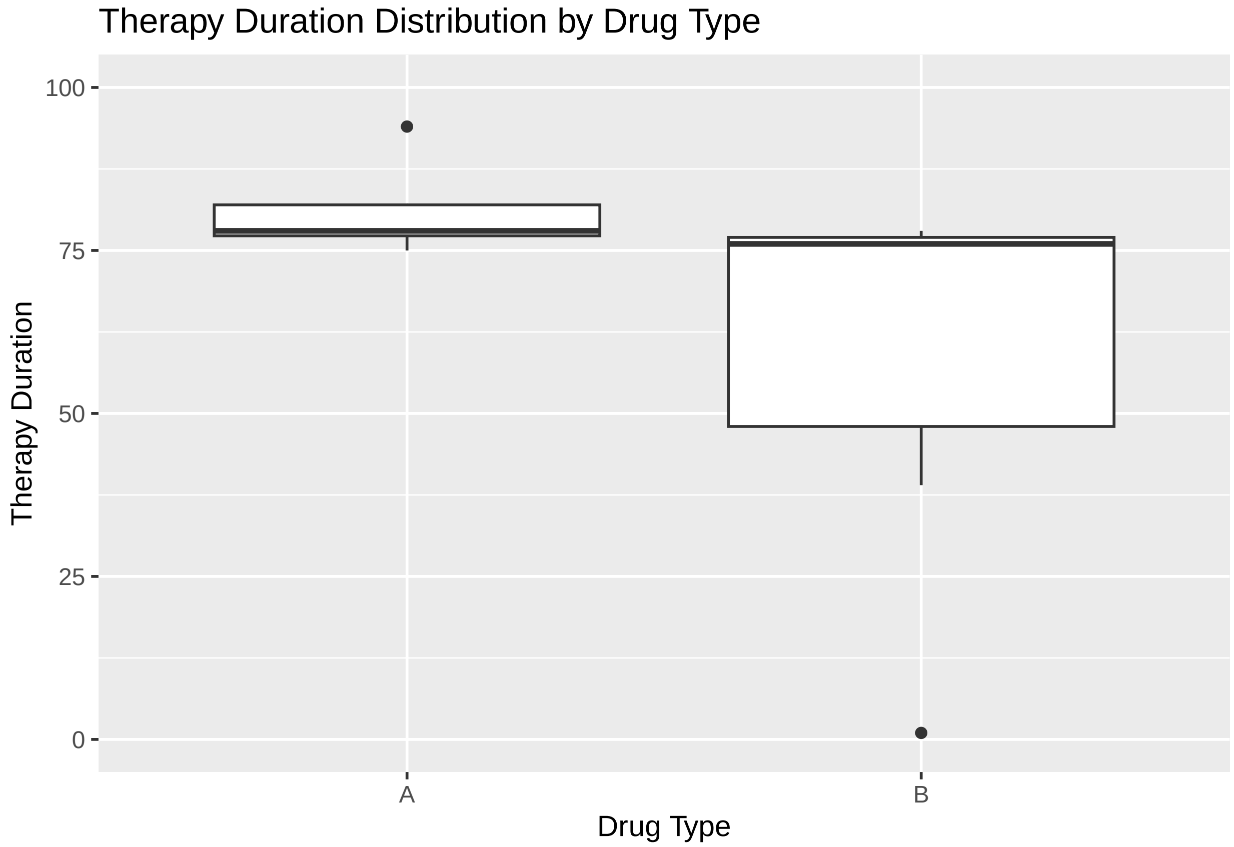
```
## `summarise()` has grouped output by 'diagnosis'. You can override using the
## `.groups` argument.
```

```r
frequency <-frequency[frequency$drug != "",]

# 4. Findings:
print(frequency)
```

```
## # A tibble: 8 x 3
## # Groups:   diagnosis [2]
##   diagnosis     drug  frequency
##   <chr>         <chr>     <int>
## 1 Breast Cancer AB           18
## 2 Breast Cancer B             7
## 3 Breast Cancer C             6
## 4 Breast Cancer A             4
## 5 Colon Cancer  AB            4
## 6 Colon Cancer  B             4
## 7 Colon Cancer  C             4
## 8 Colon Cancer  D             4
```

```r
# Breast Cancer: Based on the frequency table, the combination therapy of A and B and
# monotherapy B, C, A are used for patients. The combination therapy of treatment A and
# B is indicated as the most used first-line treatment with the highest frequency followed
# by Monotherapy B, C, A
# Colon Cancer: Based on the frequency table, the combination therapy of A and B and
# monotherapy B, C, D are used for patients. However, there is no treatment standing
# out as the most-used first-line treatment. All four types of treatment have the same frequency.



#Question 4 ------------------------------------
# 1. Filter data to only keep patients who have breast cancer treated with Monotherapy
breast_mo <- merge_first[merge_first$diagnosis == "Breast Cancer" & merge_first$drug %in% c("A","B"),]

# 2. Extract the patient id and treatment history
bmp <- breast_mo$patient_id
bmp_trt <- trt[trt$patient_id %in% bmp,]

# 3. Derive the metric of "Duration of Therapy"
bmp_dur <- bmp_trt %>%
    group_by(patient_id) %>%
    summarise(first_date = min(daten),
              last_date = max(daten)) %>%
    mutate(duration = as.numeric(difftime(last_date, first_date, units="days")) + 1) %>%
    mutate(patient_id = as.character(patient_id))

breast_mo <- breast_mo %>%
    left_join(bmp_dur, by='patient_id')
```

```r
# 4. Visualization for the Distribution of Duration of Therapy
# In order to better present the data in the histogram, the outlier are excluded from the plot.
breast_p <- breast_mo[breast_mo$duration < 1000,]
ggplot(breast_p, aes(x = drug, y = duration)) +
  geom_boxplot() +
  labs(x = "Drug Type", y = "Therapy Duration") +
  ggtitle("Therapy Duration Distribution by Drug Type") +
  scale_y_continuous(limits = c(0, 100))
```



```r
# 5. Perform Statistical Test
# Assumption: The Duration of Therapy follows Normal Distribution
# For this question, since we want to investigate if the duration of therapy are
# different for two types of Monotherapy of A and B, we choose the two-sample t-test
# to determine with significance level at 0.05 if two population means are different.
t_test_result <- t.test(duration ~ drug, data = breast_mo)

# 6. Extract p-value
p_value <- t_test_result$p.value
print(p_value)
```

```
## [1] 0.4417785
```

```r
# 7. Compare p-value to significance level
if (p_value < 0.05) {
  print("Reject the null hypothesis")
} else {
  print("Fail to reject the null hypothesis")
}
```

```
## [1] "Fail to reject the null hypothesis"
```

```
# 8. Conclusion
# According to our test result, our p value is 0.4417785 which is bigger than 0.05 and
# we fail to reject the null hypothesis. We can conclude that there is no significant
# variation in terms of duration of therapy between two monotherapy.
```