ASSIGNMENT 9

GSI Intro to Big Data and Data Mining

The University of Texas at Austin

Zhaowen Fan
Rafael Ignacio Gonzalez Chong

Table of Contents

(1) Save the data to a excel or CSV file and read into R for analysis. (2 points)

```
> library(readr)
> library(ggplot2)
> library(tidyr)
> #Use the following dataset to:
> #(1) Save the data to a excel or CSV file and read into R for analysis.
> #(2 points)
> datos_correct <- read_csv("datos_correct.csv")
New names:
• `3` -> `3...2`
• `5` -> `5...3`
• `3` -> `3...4`
• `5` -> `5...6`
• `3` -> `3...8`
• `4` -> `4...9`
• `4` -> `4...10`
Rows: 9 Columns: 10
── Column specification ────────────────────────────────────────────────────────
Delimiter: ","
dbl (10): 7, 3...2, 5...3, 3...4, 1, 5...6, 10, 3...8, 4...9, 4...10

ℹ Use `spec()` to retrieve the full column specification for this data.
ℹ Specify the column types or set `show_col_types = FALSE` to quiet this message.
> datos_long <- datos_correct |>
+   pivot_longer(
+     everything(),               # take columns
+     names_to = "column",        # stack
+     values_to = "duration"      # duration
+   )
```

Fig. 1 Reading data from csv file.

(2) Make a histogram of the duration of days of hospital stays. Ensure the histogram is labelled appropriately. Use a width of 1 day. Describe the shape center and spread of the data. Are there any outliers? (7 points)

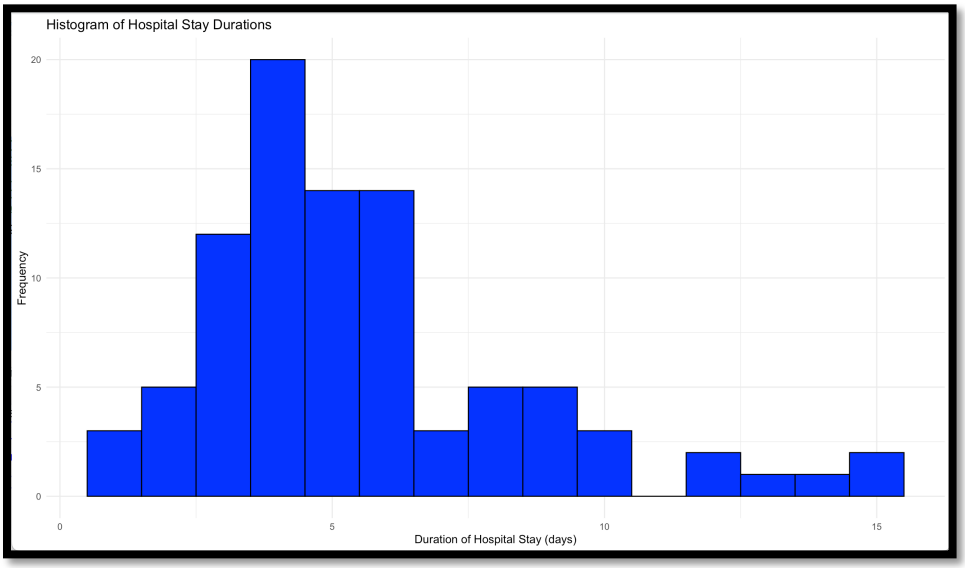

Fig. 2 Histogram of the duration of days of hospital stays.

The histogram shows the distribution of hospital stay durations. Most patients stayed between 3 and 7 days, with the highest frequency around 4 days. Short and extended stays are less common, indicating a slightly right-skewed distribution with few long hospitalizations.

(3) Find the mean, median, standard deviation, first and third quartiles, minimum and maximum of the durations of hospital stay in the sample. Summarize these values in a table that you create in EXCEL or WORD. In other words, do *not* simply copy and paste R output. Given the shape of the distribution, what is the best single number summary of the center of the distribution? What is the best single number summary of the spread of the distribution? (6 points)

```
> #(3) Find the mean, median, standard deviation, first and third quartiles,
> #minimum and maximum of the durations of hospital stay in the sample.
> #Summarize these values in a table that you create in EXCEL or WORD.
> #In other words, do *not* simply copy and paste R output.
> #Given the shape of the distribution, what is the best single number summary of
> #the center of the distribution?  What is the best single number summary of
> #the spread of the distribution? (6 points)
> valores_todos <- as.vector(as.matrix(datos_correct)) # change as a vector
> summary_table <- data.frame(
+   Statistic = c("Mean", "Median", "Standard Deviation",
+                 "1st Quartile (Q1)", "3rd Quartile (Q3)",
+                 "Minimum", "Maximum"),
+   Value = c(mean(valores_todos),
+             median(valores_todos),
+             sd(valores_todos),
+             quantile(valores_todos, 0.25),
+             quantile(valores_todos, 0.75),
+             min(valores_todos),
+             max(valores_todos))
+ )
>
> print(summary_table)
           Statistic     Value
1               Mean  5.555556
2             Median  5.000000
3 Standard Deviation  3.002288
4  1st Quartile (Q1)  4.000000
5  3rd Quartile (Q3)  6.000000
6            Minimum  1.000000
7            Maximum 15.000000
```

Fig. 3 Metrics from the Hospital Stays.

The best single number summary of the center is the median (5 days). For the spread, the interquartile range (IQR) of 2 days is best.

(4) Assume that the literature on this topic suggests that the distribution of days of hospital stay is normally distributed with a mean of 5 and a standard deviation of 3. Use R to determine the probabilities below based on the normal distribution: (a) What percentage of patients are in the hospital for less than a week? (2 points)

```
> #(4) Assume that the literature on this topic suggests that the distribution
> #of days of hospital stay are normally distributed with a mean of 5 and
> #a standard deviation of 3.  Use R to determine the probabilities below
> #based on the normal distribution: #(a) What percentage of patients are in
> #the hospital for less than a week? (2 points)
> mean_val <- 5
> sd_val <- 3
> p_less_than_7 <- pnorm(7, mean = mean_val, sd = sd_val)
> percentage <- p_less_than_7 * 100
> cat("Percentage of patients in hospital < 7 days:", round(percentage, 2), "%")
Percentage of patients in hospital < 7 days: 74.75 %
```

Fig. 4 People that stay less than a week.

(b) Recent publications have indicated that hypervirulent strains of C. Difficile are on the rise. Such strains are associated with poor outcomes, including extended hospital stays. An investigator is interested in showing that the average hospital stays durations have increased versus published literature. He has a sample of 10 patients from his hospital. If the published data are consistent with the truth, what is the probability that the sample mean in his sample will be greater than 7 days? (3 points)

```
> #(b) Recent publications have indicated that hypervirulent strains of C.
> #Difficile are on the rise.
> #Such strains are associated with poor outcomes,
> #including extended hospital stays.
> #An investigator is interested in showing that the average hospital
> #stay durations have increased versus published literature.
> #He has a sample of 10 patients from his hospital.
> #If the published data are consistent with the truth,
> #what is the probability that the sample mean in his sample will
> #be greater than 7 days? (3 points)
> mu <- 5
> sigma <- 3
> n <- 10
> se <- sigma / sqrt(n) # standard error
> p_mean_greater_7 <- pnorm(7, mean = mu, sd = se, lower.tail = FALSE) # mean > 7
> cat("Probability the sample mean > 7 days:", round(p_mean_greater_7 * 100, 2), "%")
Probability the sample mean > 7 days: 1.75 %
```

Fig. 5 Probability more than 7 days.

Appendices (Code)

```
#ASSIGNMENT 9

#GSI Intro to Big Data and Data Mining

#Zhaowen Fan

#Rafael Ignacio Gonzalez Chong



library(readr)

library(ggplot2)

library(tidyr)

#Use the following dataset to:

#(1) Save the data to a excel or CSV file and read into R for analysis.

#(2 points)

datos_correct <- read_csv("datos_correct.csv")

datos_long <- datos_correct |>

  pivot_longer(

    everything(),               # take columns

    names_to = "column",        # stack

    values_to = "duration"      # duration

  )



#(2) Make a histogram of the duration of days of hospital stays.
```

```
#Ensure the histogram is labelled appropriately.
#Use a width of 1 day.
#Describe the shape center and spread of the data.
#Are there any outliers? (7 points)
ggplot(datos_long, aes(x = duration)) +
  geom_histogram(binwidth = 1, fill = "blue", color = "black") +
  labs(
    title = "Histogram of Hospital Stay Durations",
    x = "Duration of Hospital Stay (days)",
    y = "Frequency"
  ) +
  theme_minimal()



#(3) Find the mean, median, standard deviation, first and third quartiles,
#minimum and maximum of the durations of hospital stay in the sample.
#Summarize these values in a table that you create in EXCEL or WORD.
#In other words, do *not* simply copy and paste R output.
#Given the shape of the distribution, what is the best single number summary of
#the center of the distribution?  What is the best single number summary of
#the spread of the distribution? (6 points)
valores_todos <- as.vector(as.matrix(datos_correct)) # change as a vector
summary_table <- data.frame(
```

```
  Statistic = c("Mean", "Median", "Standard Deviation",
                "1st Quartile (Q1)", "3rd Quartile (Q3)",
                "Minimum", "Maximum"),
  Value = c(mean(valores_todos),
            median(valores_todos),
            sd(valores_todos),
            quantile(valores_todos, 0.25),
            quantile(valores_todos, 0.75),
            min(valores_todos),
            max(valores_todos))
)
print(summary_table)



#(4) Assume that the literature on this topic suggests that the distribution
#of days of hospital stay are normally distributed with a mean of 5 and
#a standard deviation of 3.  Use R to determine the probabilities below
#based on the normal distribution: #(a) What percentage of patients are in
#the hospital for less than a week? (2 points)
mean_val <- 5
sd_val <- 3
p_less_than_7 <- pnorm(7, mean = mean_val, sd = sd_val)
percentage <- p_less_than_7 * 100
```

```
cat("Percentage of patients in hospital < 7 days:", round(percentage, 2), "%")



#(b) Recent publications have indicated that hypervirulent strains of C.
#Difficile are on the rise.
#Such strains are associated with poor outcomes,
#including extended hospital stays.
#An investigator is interested in showing that the average hospital
#stay durations have increased versus published literature.
#He has a sample of 10 patients from his hospital.
#If the published data are consistent with the truth,
#what is the probability that the sample mean in his sample will
#be greater than 7 days? (3 points)
mu <- 5
sigma <- 3
n <- 10
se <- sigma / sqrt(n) # standard error
p_mean_greater_7 <- pnorm(7, mean = mu, sd = se, lower.tail = FALSE) # mean > 7
cat("Probability the sample mean > 7 days:", round(p_mean_greater_7 * 100, 2), "%")
```