Jiankun_Dong_HW1

Jiankun (Bob) Dong

2023-09-18

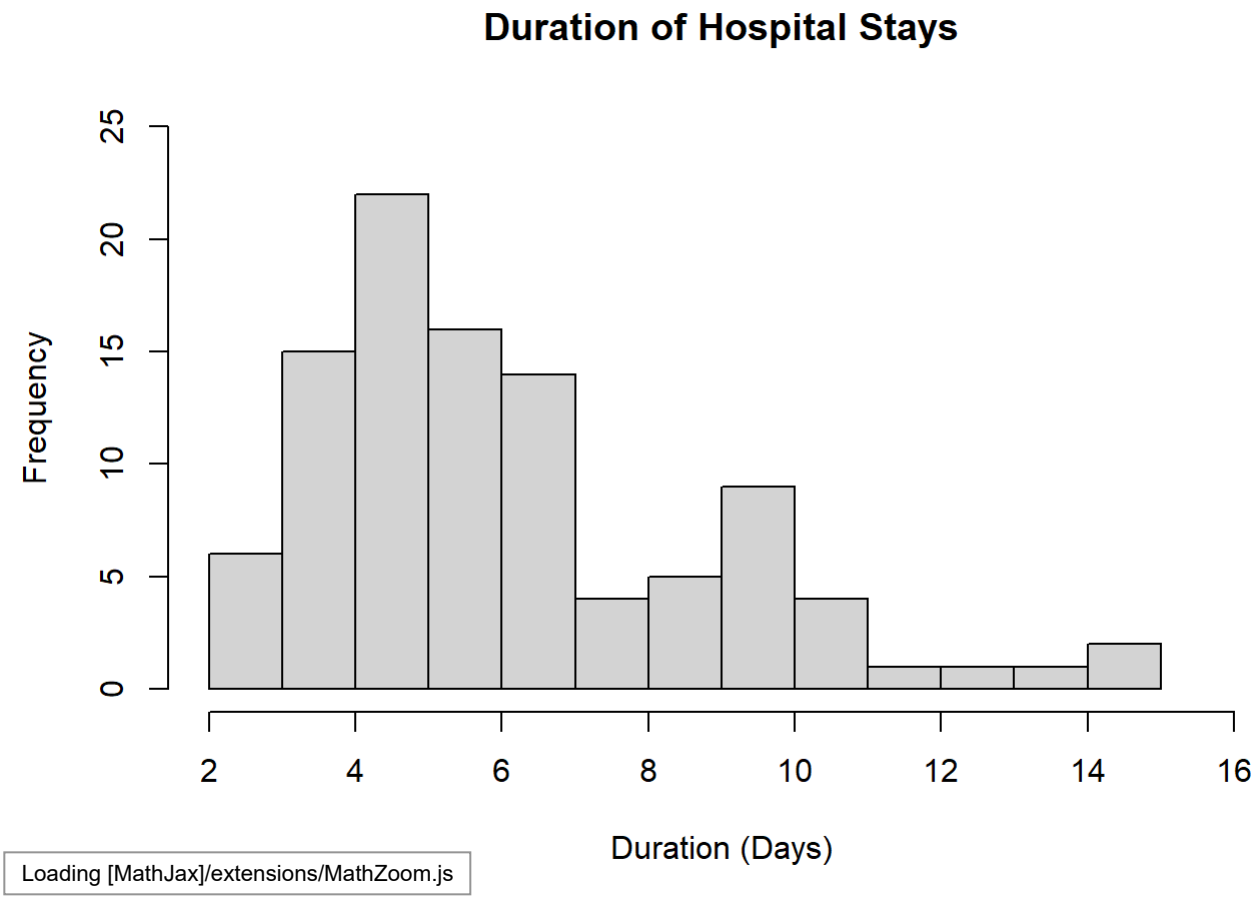
Using R Markdown to generate the file, therefore the r code are in-line.

Problem 1: Using the given dataset

```
#Load kable lib for generating tables
library(kableExtra)
#Loading the file
setwd("C:/BU/CSSE/CS555/HW1")
DaysRAW <- read.csv("./A01.csv",header = TRUE)
```

1. Loaded the csv file as DaysRaw.
2. Drawing the histogram based on the loaded data.

```
DaysPlot <- hist(DaysRAW$Days, main = "Duration of Hospital Stays",
                 xlab = "Duration (Days)",
                 breaks = seq(min(DaysRAW),max(DaysRAW),1),
                 xlim = c(min(DaysRAW),max(DaysRAW)+1),
                 ylim = (c(0,25)), plot = TRUE, right = F)
```



Shape: The data is right skewed.
Center: The center of the data is 5 days.
Spread: The first qualtile of the data is 4 days, and the third quantile is 7 days. With a standard diviation of 2.74379.
Outliers:
Because the iqr of the data is 3 day, we get the lower bound -0.5 and upper bound 11.5
The outliers are: 14, 13, 15, 12

3)

```
daysFrame <- data.frame(
  Mean = daysMean,
  Median = daysMedian,
  SD = daysSD,
  First_Quantile = quantile(DaysRAW$Days,.25)[[1]],
  Third_Quantile = quantile(DaysRAW$Days,.75)[[1]],
  Min = daysMin,
  Max = daysMax
)
daysTable <- kable(daysFrame,"simple")
```

Mean	Median	SD	First_Quantile	Third_Quantile	Min	Max
5.63	5	2.74379	4	7	2	15

Because the outliers are all beyond the third quantile, and the histogram is right skewed, the best value to summarize the center of this distribution is the median 5 days.

The best number to describe the spread of the data is the standard deviation 2.74379.

Problem 2:
4)
part a:

```
# Question 2 ------------------------------------------------------------
LessThanTen <- (pnorm(10,5,3) - pnorm(0,5,3))/pnorm(0,5,3,lower.tail = FALSE)
```

The percentage of the patients in hospital for less than 10 days is 94.9811103%.

part b:

```
n <- 35
SE <- 3/sqrt(n)
MoreThanSix <- pnorm(6,5,SE,lower.tail = FALSE)/(1-pnorm(0,5,SE))
```

Because the sample size 35 is larger than 30, we can use the CLT. The probability of the average of the data set being more than 6 days is 0.0243033.