Xiang Ji hw1

1. Save the data to a CSV file and read into R for analysis.

setwd("/Users/estherji")

hos <- read.csv("hospital.csv", header=FALSE)

1. 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?

Chart, histogram

Description automatically generated

The shape of this graph is Bimodal. It has two peaks. The highest peak is when the number of days of hospitalization is 4. There is a second high point when the number of days of hospitalization is 9. But the first peak is higher than the second peak.

center and spread of the data:

the range is 13, **Mean is 5.63, Median is 5, Standard deviation**

is 2.74379.

12, 13, 14, 15 seem to be outliers

（3）Find the mean, median, standard deviation, first and third quartiles, minimum and minimum 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, create your own table and describe the values.

**Mean is 5.63, Median is 5, Standard deviation**

is 2.74379.

first quartiles： 4.00

third quartiles：7.00

minimum：2.00

maximum：15.00

Graphical user interface, text, application

Description automatically generated

- Given the shape of the distribution, what is the best single number summary of the center of the distribution?

Median 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

Range is the best single number summary of the spread of the distribution.

(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 10 days? (4 points)

95.22096% of patients are in the hospital for less than 10 days.

(b) What percentage of patients are in the hospital between 3 and 10 days?

69.97171% of patients are in the hospital between 3 and 10 days.

R code:

library(UsingR)

setwd("/Users/estherji")

#Part 1

hos <- read.csv("hospital.csv", header=FALSE)

head(hos)

# 2

typeof(hos)

a <- as.numeric(as.matrix(hos))

a

table(a)

hist(a, ylim=c(0,25),breaks = seq(0,17,1), xlab = "Number of days in hospital", main="Number of days in hospital distribution")

# 3

data <- c(mean(a), median(a), sd(a), quantile(a,0.25), quantile(a,0.75), min(a), max(a) )

datas <- matrix(data, nrow = 1, ncol = 7, byrow = TRUE)

dimnames(datas) <- list(

c("durations of hospital "),

c("mean", "median", "standard deviation", "first quartiles", "third quartiles", "minimum", "maximum"))

write.csv(datas,"hospital.csv")

#4

#(a) What percentage of patients are in the hospital for less than 10 days

pnorm(10, mean = 5, sd = 3)

#(b) What percentage of patients are in the hospital between 3 and 10 days?

b <- pnorm(10, mean = 5, sd = 3) - pnorm(3, mean = 5, sd = 3)

b