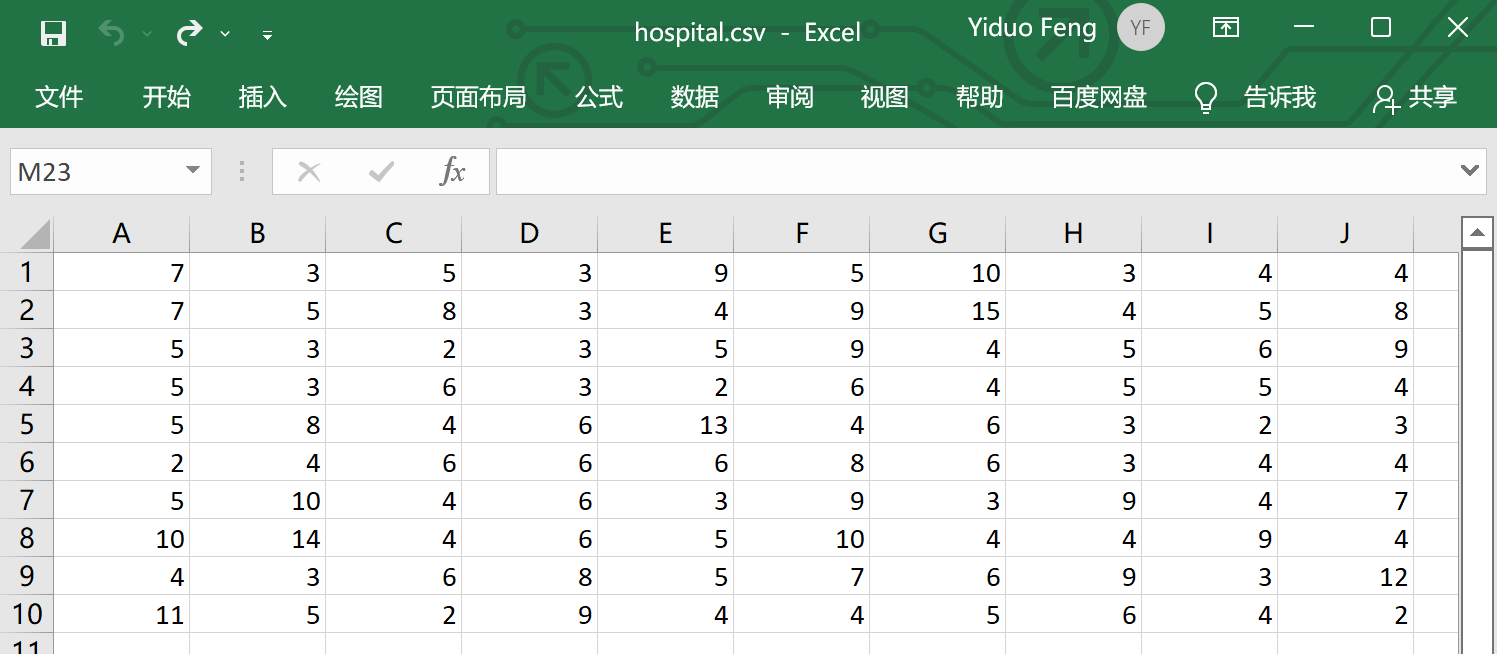
Yiduo Feng

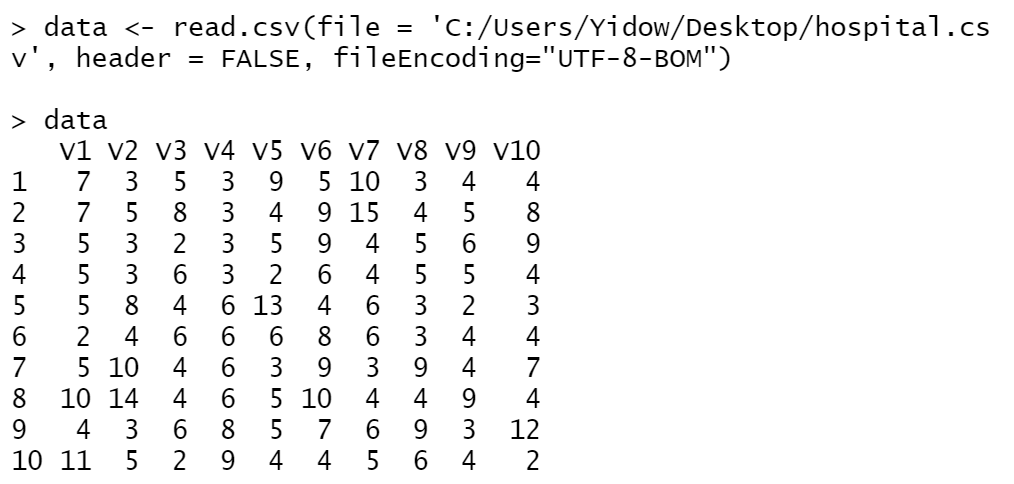
CS 555

Homework 1

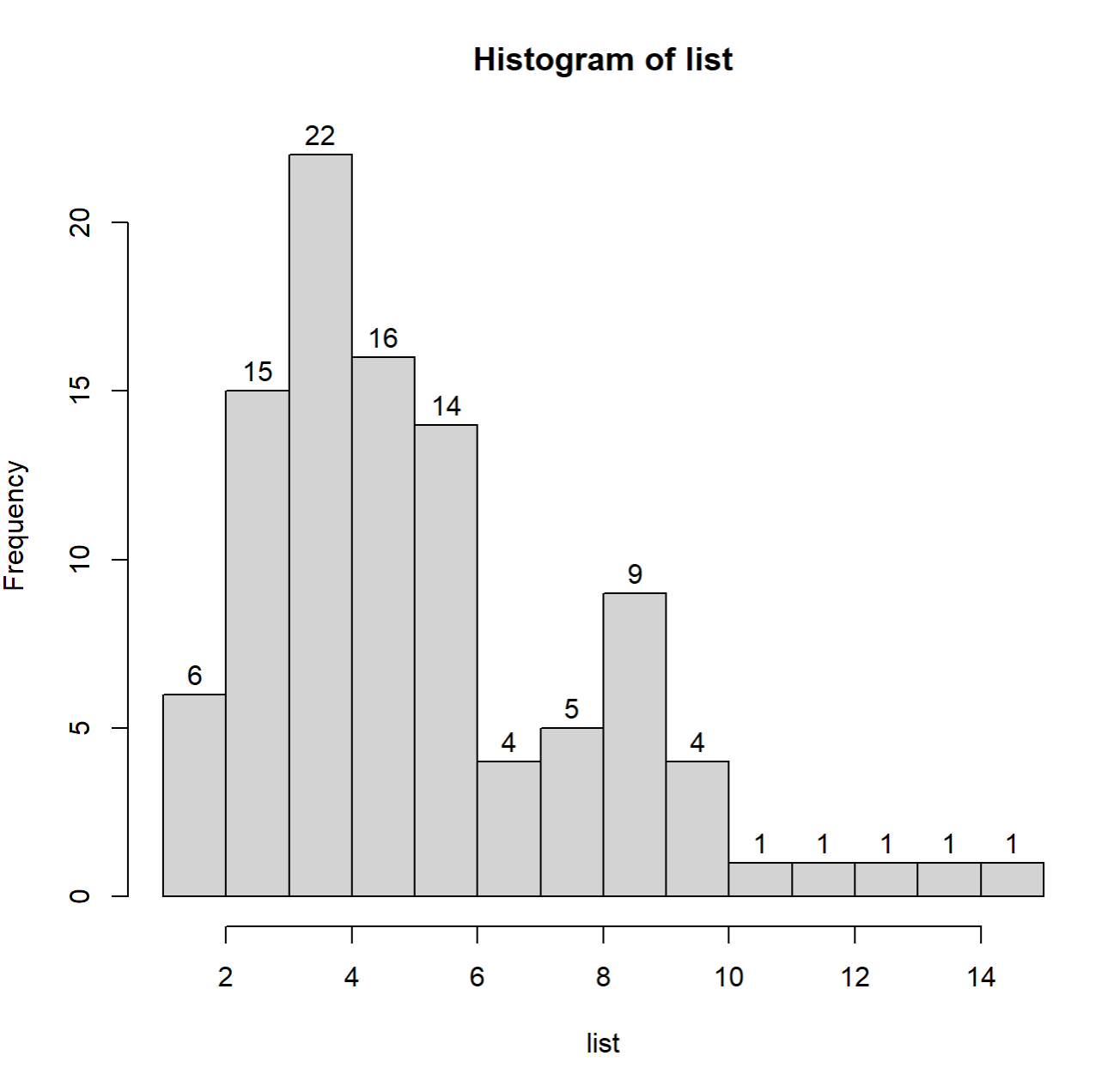
07/12/2022

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

****

****

**(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?**



According to the graph, it’s right skewed, so we use median as the center which is 5.

IQR is the best measure of spread which is Q3 -Q1 = 7-4 = 3.

Q1-1.5IQR = -0.5, Q3+1.5\*IQR = 11.5

Yes. There are some outliers which is 12, 13, 14, 15.

**(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. You should be reporting a nicely labeled and formatted table.**

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

Median 5

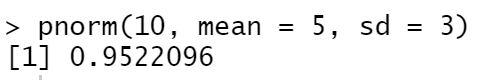
**- What is the best single number summary of the spread of the distribution?**

IQR = Q3 -Q1 = 7-4 = 3

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Min | 1st Q | Median | Mean | 3rd Q | Max | standard deviation |
| 2 | 4 | 5 | 5.63 | 7 | 15 | 2.74379 |

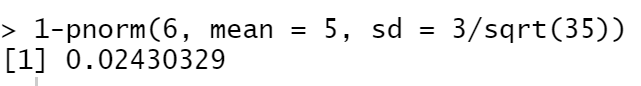
**(4) Assume that the literature on this topic suggests that the distribution of days of hospital stay are normally distributed with a mean of 5 and a standard deviation of 3. Use R to determine the probabilities below based on the normal distribution described above (you should not be using the data set given on the following page):**

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



Thus, 95.22% of patients are in the hospital for less than 10 days.

**(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 duration have increased versus published literature. He has a sample of 35 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 6 days?**



Thus, 2.43% that the sample mean in his sample will be greater than 6 days.

**Code:**

data <- read.csv(file = 'C:/Users/Yidow/Desktop/hospital.csv', header = FALSE, fileEncoding="UTF-8-BOM")

data

list <- c()

for (i in 1:ncol(data)){

list <- c(list,data[,i])

}

list

h <- hist(list, breaks=seq(1,15))

text(h$mids,h$counts,labels=h$counts, adj=c(0.5, -0.5))

h

summary(list)

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

1-pnorm(6, mean = 5, sd = 3/sqrt(35))