MET CS555: Homework 1

Patrick Ryan

(7/9/2018)

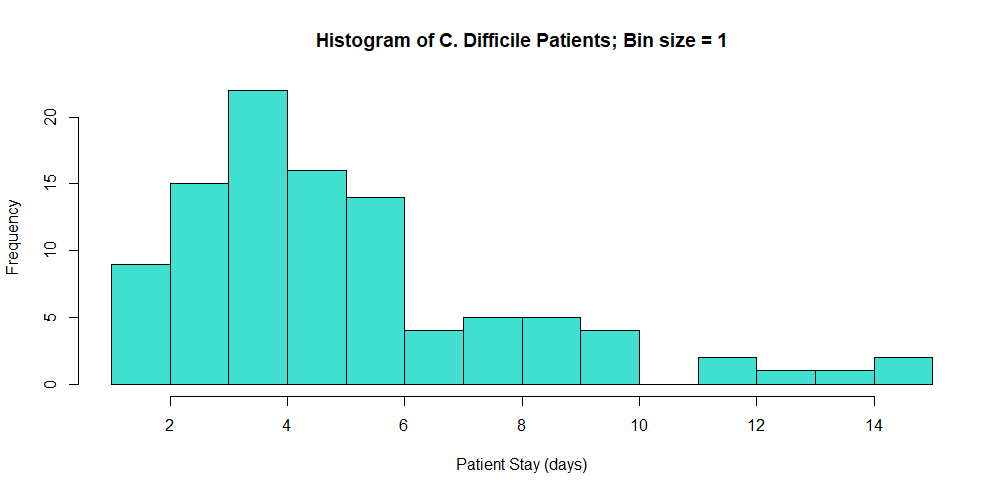


(1) Save the data to an excel or CSV file and read into R for analysis. (2 points)

|  |
| --- |
| df = read.csv("hospital\_admittance.csv", header=F, sep=",", fileEncoding="UTF-8-BOM", stringsAsFactors = F)  df  #Reassign the list as a vector  df <- as.vector(unlist(df)) |

(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)

[Figure 1: Histogram of Patients]



Description of Patient Data Shape:

* Data is very spread and skews heavily right (Skewness coefficient of 1.23)
* Data is binned with size/width of 1 day.
* Data is fairly well centered on its mean (5)
* Significant statistical outliers exist (10, 15, 10, 13, 14, 10, 10, 15, 12, 12)

(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)

[Figure 2: Table of Statistical Metrics for Patient Data]

|  |  |
| --- | --- |
| **Metric** | **Value** |
| Mean | 5.45 |
| Median | 5 |
| Standard Deviation | 2.962305 |
| First Quartile | 1 |
| Third Quartile | 5 |
| Minimum | 1 |
| Maximum | 15 |

Given the shape of the distribution:

* The best single number summary of the center of the distribution is the median because it is less influenced by statistical outliers.
* The best single number summary of the spread of the distribution is the standard deviation, because it shows the degree to which the population deviates from its mean.

(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)

**74.75075 %**

(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)

**7.827011 %**

[Figure 2: R Code Used to Generate Answers for Homework 1]

|  |
| --- |
| library(ggplot2)  library(e1071)  setwd("C:/Users/patry/OneDrive/Desktop")  #1)  df = read.csv("hospital\_admittance.csv", header=F, sep=",", fileEncoding="UTF-8-BOM", stringsAsFactors = F)  df  #Reassign the list as a vector  df <- as.vector(unlist(df))  df  #2)  # Histogram  hist(df, xlab="Patient Stay (days)", col="turquoise", main = ("Histogram of C. Difficile Patients; Bin size = 1"), breaks=seq(max(df)))  #3)  mean(df)  median(df)  sd(df)  quantile(df)[1]  quantile(df)[3]  min(df)  max(df)  # Outliers  boxplot(df, horizontal = T, col="turquoise")$out  skewness(df)  #4)  # a)  # Distribution of C. Difficle stays less than a week  pnorm(7, mean = 5, sd = 3, lower.tail = T)  # (b)  sample.pop = 10  sample.sd = 3  sample.var <- sample.sd/sqrt(sample.pop)  sample.var  x <- (5 - 2) / sample.var #(3.162278) sd above the mean  percentage <- 100 \* pnorm(x, lower.tail=F)  percentage |