CS555, Data Analysis and Visualization Homework 1

Jefferson Parker, [japarker@bu.edu](mailto:japarker@bu.edu)

See the accompanying R code at the end of this document.

1. The *C. difficile* data file was saved to a local directory as a .csv file and read into R as:

inputdir <- "C:/Users/jparker/Code/Input";

setwd(inputdir);

difi.data <- read.csv(file = "CdifficileData.csv", header = FALSE);

1. The histogram of hospital stay duration is shown below. The data is **right skewed** with the center of the frequency distribution to the left, **centered at a value of four days**, and a long tail to the right with a **range of 1 – 15 days**.



There are several outliers in the data set, defined as values less than Q1 – (1.5 \* IQR) or Q3 + (1.5 \* IQR), where IQR is the InterQuartile Range. The **outliers** are at:

|  |  |
| --- | --- |
| **Value** | **Frequency** |
| 1 | 4 |
| 9 | 5 |
| 10 | 4 |
| 12 | 2 |
| 13 | 1 |
| 14 | 1 |
| 15 | 2 |

1. The summary values and standard deviation of the data set are:

|  |  |
| --- | --- |
| **Metric** | **Value (Hospital Stay in Days)** |
| Mean | 5.45 |
| Median | 5 |
| Standard Deviation | 2.96 |
| First Quartile | 4 |
| Third Quartile | 6 |
| Minimum | 1 |
| Maximum | 15 |

Given the summary values above, and considering the shape of the data in question 2, the best descriptor of the center of the data is the **median** of 5 while the best descriptor for the spread of the data is the **interquartile range** (IQR) 2. These are the best descriptors for the data because the data is skewed with known outliers, which can have outsized effect on the mean and standard deviation.

1. Assuming the literature report of hospital duration being normally distributed with a mean = 5 and standard deviation = 3:
   1. The percentage of patients in hospital for less than one week (7 days) is **74.8%**.
   2. For the hyper virulent *C. difficile* strain under investigation, the probability that the sample mean in the sample of n = 10 is greater than 7 days is **1.7%**, assuming the published data are consistent with the truth.

This is determined by first calculating the standard error of the mean, , and substituting that value for the standard deviation of our normal distribution.

**R Code**

# CS555 Data Analysis and Visualization

# Homework1.R

# Jefferson Parker, japarker@bu.edu

# 20180703

# 1. Save the data to a .csv file and read it into R.

# The file was saved locally as CdifficileData.csv.

inputdir <- "C:/Users/jparker/Code/Input";

setwd(inputdir);

difi.data <- read.csv(file = "CdifficileData.csv", header = FALSE);

# The read.csv function automatically creates a data frame.

# Convert the data to an unnamed integer vector for convenience.

difi.data <- unname(unlist(difi.data));

# 2. Make a histogram of the duration of hospital days, binned by single

# days. Lable the plot apporpriately.

hist(difi.data,

breaks = 14,

xlim = c(0,15),

main = "Distribution of C. dificile Hospital Stay Duration",

xlab = "Days",

ylab = "Frequency"

);

# Determine if there are outliers.

find.outliers <- function(x){

fiveNum <- fivenum(x);

firstQuartile <- fiveNum[2];

thirdQuartile <- fiveNum[4];

iqrVal <- fiveNum[4] - fiveNum[2];

upperBound <- thirdQuartile + (1.5 \* iqrVal);

lowerBound <- firstQuartile - (1.5 \* iqrVal);

return(sort(x[x <= lowerBound | x >= upperBound]));

}

table(find.outliers(difi.data));

# 3. Find the mean, median and standard deviation, first and third

# quartiles, min, max.

summary(difi.data);

sd(difi.data);

# 4. Assume normal distribution with mean = 5 and sd = 3.

# a) What percentage of patients are in the hospital less than a week?

pnorm(7, mean = 5, sd = 3);

# b) For a sample n = 10, what is the probability a hospital stay is

# more than 7 days?

#

# First, we use the provided information to find the Standard Error of the mean.

# Then we use this as the estimate of the population standard deviation to

# calculate our probability.

se.xbar = 3/sqrt(10);

1 - pnorm(7, mean = 5, sd = se.xbar);