# (1. )

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

A:

setwd("C:/Users/XuWei/OneDrive/BU - Boston University/04 - 2022 Spring Semester/MET CS 555 A2 Data Analysis and Visualization with R/Assignment 1")

data <- read.csv("assignmentData.csv", header = FALSE)

# (2. )

Q: 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? (5 points)

A:

# shape:

# As we can see from out plot, there are more data on the left side rather than the right side.It is Skewed right.

# center:

# Since the graph is skewed right, we will use the median instead the center, which is 5.

m <- median(data1$unlist.data)

# spread:

# as we can see from the plot, we can see most of data are concentrated at a median, which means the data are not too scattered.

图表, 直方图

描述已自动生成

# Outliers:

# Yes, there are some outliers.

# Outliers: 14, 13, 15, 12

out <- boxplot.stats(data1$unlist.data.)$out

out\_ind <- which (data1$unlist.data. %in% c(out))

out\_ind

boxplot(data1$unlist.data., ylab = "values", main = "Boxplot of Duration of days of hospital stays")

mtext(paste("Outliers: ", paste(out, collapse = ",")))

图表, 箱线图

描述已自动生成

# (3. )

Q:

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, create your own table and describe the values.

- 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? (5 points)

A:

# mean: 5.63

# median: 5.00

# standard deviation: 2.74379

# first quartiles: 4.00

# third quartiles: 7.00

# minimum: 2.00

# maximum: 15.00

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
|  | Mean | Median | Standard deviation | First quartiles | Third quartiles | Minimum | Maximum |
| Number of days patient stays hospital | 5.63 | 5.00 | 2.74379 | 4.00 | 7.00 | 2.00 | 15.00 |

# standard deviation: 2.74379, Standard deviation is the best single number summary of spread of the distribution.

#(4. )

Q:

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:

1. What percentage of patients are in the hospital for less than 10 days? (4 points)

pnorm(10, 5, 3)

0.9522096

1. What percentage of patients are in the hospital between 3 and 10 days? (4 points)

pnorm(10, 5, 3) - pnorm(3, 5, 3)

0.6997171

R code:

# (1. )

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

setwd("C:/Users/XuWei/OneDrive/BU - Boston University/04 - 2022 Spring Semester/MET CS 555 A2 Data Analysis and Visualization with R/Assignment 1")

data <- read.csv("assignmentData.csv", header = FALSE)

# (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?

# make all the data become in one column

data1 <- data.frame(unlist(data))

plot(hist(data1$unlist.data, w=1), xlab = "Duration of Days of Hospital Htays", ylab = "Counts", main = "Histogram of Duration", col = "grey")

# shape:

# As we can see from out plot, there are more data on the left side rather than the right side.It is Skewed right.

# center:

# Since the graph is skewed right, we will use the median instead the center, which is 5.

m <- median(data1$unlist.data)

# spread:

# as we can see from the plot, we can see most of data are concentrated at a median, which means the data are not too scattered.

# Outliers:

# Yes, there are some outliers.

# Outliers: 14, 13, 15, 12

out <- boxplot.stats(data1$unlist.data.)$out

out\_ind <- which (data1$unlist.data. %in% c(out))

out\_ind

boxplot(data1$unlist.data., ylab = "values", main = "Boxplot of Duration of days of hospital stays")

mtext(paste("Outliers: ", paste(out, collapse = ",")))

# (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\* simple copy and paste output. create your own table and describe the values.

summary(data1$unlist.data.)

sd(data1$unlist.data.)

# mean: 5.63

# median: 5.00

# standard deviation: 2.74379

# first quartiles: 4.00

# third quartiles: 7.00

# minimum: 2.00

# maximum: 15.00

# - 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?

# standard deviation: 2.74379, Standard deviation is the best single number summary of spread of the distribution.

#(4. )

# Assume that the literature on his 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?

pnorm(10, 5, 3)

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

pnorm(10, 5, 3) - pnorm(3, 5, 3)