Dhakal_module1

Kokil Dhakal

2023-11-10

Question1

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

Answer:

```
library(readxl)
module1.data <- read_xlsx("/Users/kokildhakal/Desktop/STUDY/BU/9.CS555/Module1/data.xlsx",
                          col_names = FALSE)
```

```
## New names:
## * `` -> `...1`
## * `` -> `...2`
## * `` -> `...3`
## * `` -> `...4`
## * `` -> `...5`
## * `` -> `...6`
## * `` -> `...7`
## * `` -> `...8`
## * `` -> `...9`
## * `` -> `...10`
```

```
head(module1.data)
```

```
## # A tibble: 6 x 10
##    ...1  ...2  ...3  ...4  ...5  ...6  ...7  ...8  ...9 ...10
##   <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl> <dbl>
## 1     7     3     5     3     9     5    10     3     4     4
## 2     7     5     8     3     4     9    15     4     5     8
## 3     5     3     2     3     5     9     4     5     6     9
## 4     5     3     6     3     2     6     4     5     5     4
## 5     5     8     4     6    13     4     6     3     2     3
## 6     2     4     6     6     6     8     6     3     4     4
```

In this section, I copied and paste the dataset to the excel file and then used read_xlsx method to read the excel file. also,I used col_names = false as parameter as we do not have any column name in the dataset.

Question2

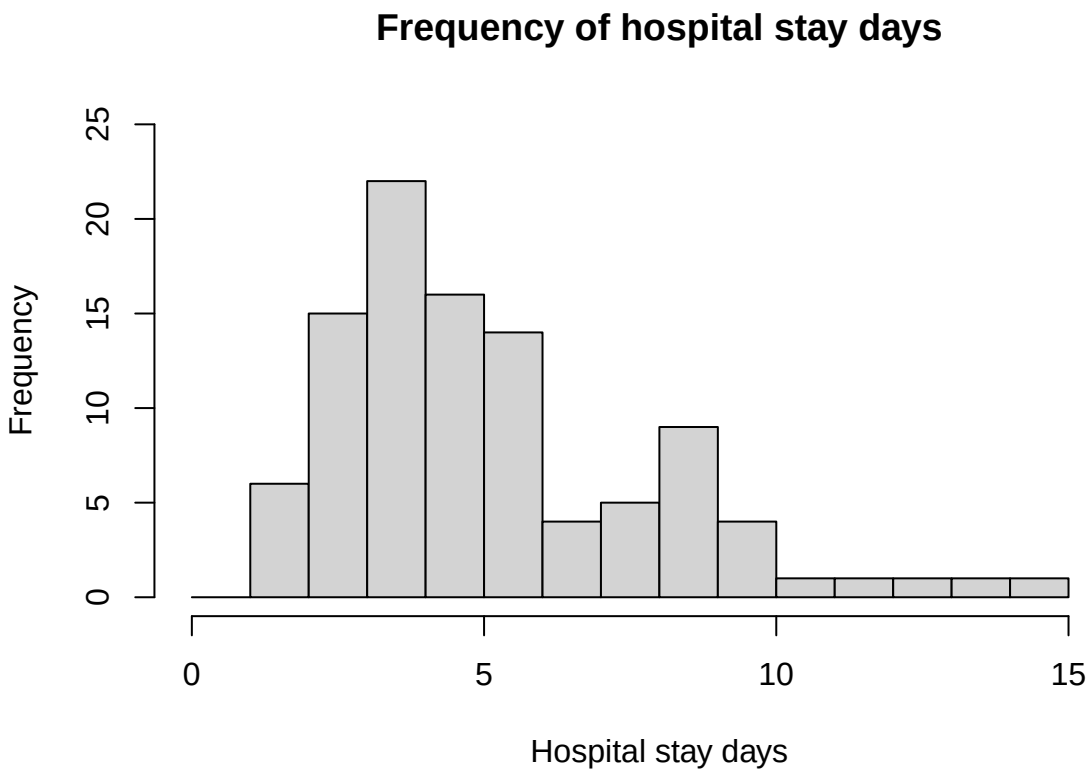
2.Make a histogram of the duration of days of hospital stays. Ensure the histogram is labeled appropriately and add the screenshot here. Use a width of 1 day. Describe the shape, center, and spread of the data. Are there any outliers?

Answer:

```
days.vectors <- unlist(module1.data,use.names = FALSE)
days.vectors
```

```
##   [1]  7  7  5  5  5  2  5 10  4 11  3  5  3  3  8  4 10 14  3  5  5  8  2  6  4
##  [26]  6  4  4  6  2  3  3  3  3  6  6  6  6  8  9  9  4  5  2 13  6  3  5  5  4
##  [51]  5  9  9  6  4  8  9 10  7  4 10 15  4  4  6  6  3  4  6  5  3  4  5  5  3
##  [76]  3  9  4  9  6  4  5  6  5  2  4  4  9  3  4  4  8  9  4  3  4  7  4 12  2
```

```
hist(days.vectors,breaks = seq(0,max(days.vectors),1),
     xlab ="Hospital stay days",
     main = "Frequency of hospital stay days",
     ylim = c(0,25),
     xlim = c(0,16))
```




```r
mean(days.vectors)
```

```
## [1] 5.63
```

```r
median(days.vectors)
```

```
## [1] 5
```

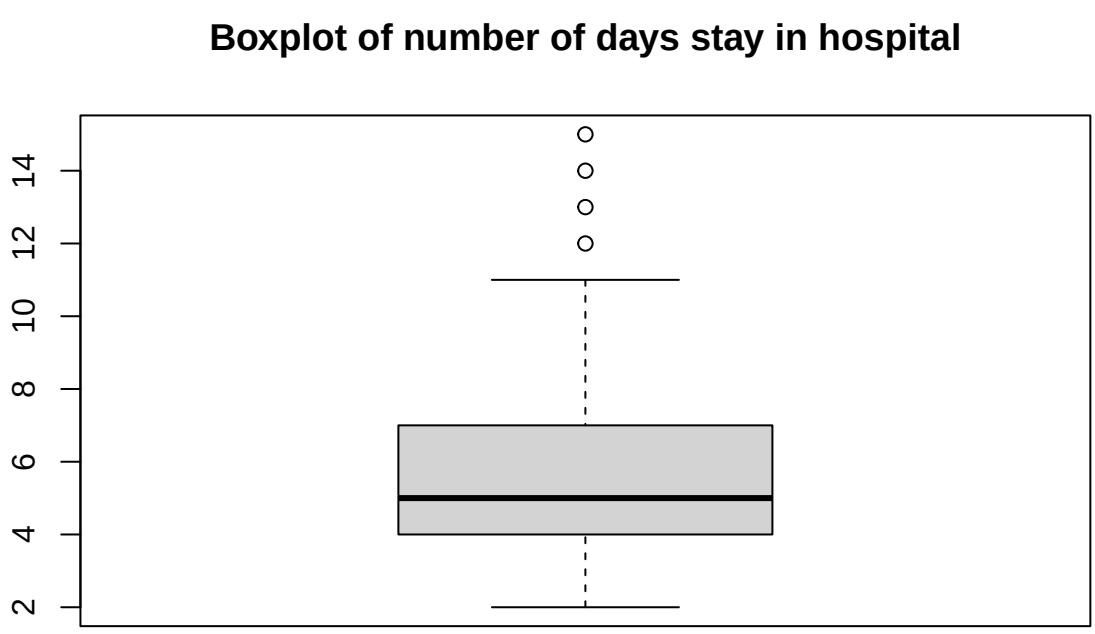
```r
sd(days.vectors)
```

```
## [1] 2.74379
```

```r
IQR(days.vectors)
```

```
## [1] 3
```

```r
boxplot(days.vectors,main="Boxplot of number of days stay in hospital")
```




```
table(days.vectors)
```

```
## days.vectors
##  2  3  4  5  6  7  8  9 10 11 12 13 14 15
##  6 15 22 16 14  4  5  9  4  1  1  1  1  1
```

Since, plotting histogram need vectors not the data frame. I need to convert dataframe to vectors using unlist method.

Shape of data: from above histogram, data does not look symmetrical, and it is right skewed.This data set does not have values less than 2 that is why we do not see bar for bin 0 to 1.Each bar of this histogram represent 1 day.

Center of data: for the center of data, I am calculating Mean and median. mean is 5.63 and median is 5 for the given data.Mode is day which has high number of patients stays which is 4.This means there are 22 patients who stays 4 days in the hospital

Spread of data: Spread of data is usually determined by Standard deviation and interquartile range. Here, Standard deviation is 2.74379 and Inter Quartile Range: 3. And finally, as from the boxplot there are four outliers can be seen.This boxplot tells that there are outliers present in the given dataset.

Question 3

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?

- What is the best summary evaluation number for the distribution spread? (note: this is skewed distribution)

Answer:

```
summary.quest3 <- data.frame(
  mean=mean(days.vectors),
  median=median(days.vectors),
  sd=round(sd(days.vectors),2),
  min=min(days.vectors),
  max=max(days.vectors),
  Q1=quantile(days.vectors,probs = 0.25),
  Q3=quantile(days.vectors,probs = 0.75),
  row.names = "data1"

)
summary_table <- kable(summary.quest3,format = "simple",row.names = FALSE,
                       align = "c",caption = "A table of values for given data")
summary_table
```

Table 1: A table of values for given data

mean	median	sd	min	max	Q1	Q3
5.63	5	2.74	2	15	4	7

I made a dataframe which contain mean, median sd,min,max Q1 and Q3 of given data.By using package KableExtra, I made a table which has name and their respective values.

The median is a robust measure of central of distribution that is less influenced by extreme values or outliers compared to the mean, which can be heavily affected by skewed tails. The median of given data is 5.In our case, the dataset has outliers and is skewed so median is the best measure of central of distribution.

For skewed data, the best summary evaluation number for distribution spread is often the Interquartile Range (IQR). The IQR is a robust measure of spread that is less sensitive to outliers and works well for both symmetric and skewed distributions. Interquartile Range of given data is Q3-Q1 i.e., 7-4=3.

Question4

4.Assume that the literature on this topic suggests that the distribution of days of hospital stay is 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?

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 has 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?

```r
#a.
mean.val <- 5
sd.val <- 3
q.val <- 10
probablity_10 <- pnorm(q.val,mean = 5,sd = 3)
cat("There are approximately",round(probablity_10*100,2),"% of patients are
      in the hospital for less than 10 days")
```

```
## There are approximately 95.22 % of patients are
##         in the hospital for less than 10 days
```

In this problem, we have given mean and sd of the days of hospital stays of patients. we need to find the percentage of patients are in the hospital for less than 10 days. Since, distribution type is normal distribution and q-value is 10 we can simply use pnorm method for normally distributed dataset to calculate the probablity and then convert that probablity to percentage by multiplying with 100.

```
#b
population.mean <- 5
population.sd <- 3
sample.size=35
sample.mean <- 6
#finding standard error.
standard.error <- population.sd/sqrt(sample.size)

#finding  z-score
zscore <-  (sample.mean - population.mean)/standard.error
#now finding the provability.since, it is greater than 6 days
probality_6 <- pnorm(zscore,lower.tail = FALSE) # or 1-pnorm(zscore)
cat("The probablity that the sample mean in investigator's sample will be
      greater than 6 days is",probality_6," or approximately",
      round(probality_6*100,2),"%")
```

```
## The probablity that the sample mean in investigator's sample will be
##       greater than 6 days is 0.02430329  or approximately 2.43 %
```

In this problem I will be applying central limit theorem which states that for sufficient sample size, sample follows the normal distribution regardless of the type of population distribution. For this question I will be finding standard error first and then finding z-score and then calculating pnorm value as shown in code above.