|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete |
| Results of rolling a dice | Discrete |
| Weight of a person | Continuous |
| Weight of Gold | Continuous |
| Distance between two places | Continuous |
| Length of a leaf | Continuous |
| Dog's weight | Continuous |
| Blue Color | Discrete |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Discrete |

Q1) Identify the Data type for the Following:

Q2) Identify the Data types, which were among the following

Nominal, Ordinal, Interval, Ratio.

|  |  |
| --- | --- |
| Data | Data Type |
| Gender | Nominal |
| High School Class Ranking | Ordinal |
| Celsius Temperature | Interval |
| Weight | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Nominal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| Level of Agreement | Nominal |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Interval |
| Blood Group | Nominal |
| Time Of Day | Ordinal |
| Time on a Clock with Hands | Ordinal |
| Number of Children | Nominal |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Ordinal |
| Years of Education | Ordinal |

Q3) Three Coins are tossed, find the probability that two heads and one tail are obtained?

Ans.) 3/8

Q4) Two Dice are rolled, find the probability that sum is

1. Equal to 1
2. Less than or equal to 4
3. Sum is divisible by 2and 3

ANS) a) 0

b) 1/6

c) (1,5),(3,3),(2,4),(4,2),(5,1), (6,6)

ans:- 1/6

Q5) A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?

ANS.) 10/21

Q6) Calculate the Expected number of candies for a randomly selected child

Below are the probabilities of count of candies for children(ignoring the nature of the child-Generalized view)

|  |  |  |
| --- | --- | --- |
| CHILD | Candies count | Probability |
| A | 1 | 0.015 |
| B | 4 | 0.20 |
| C | 3 | 0.65 |
| D | 5 | 0.005 |
| E | 6 | 0.01 |
| F | 2 | 0.120 |

Child A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20

**(1\*0.015) + (4\*0.20) + (3\*0.65) + (5\*0.005) + (6\*0.01) + (2\*0.120) = 3.09**

Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points,Score,Weigh>

Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

df<-read.csv("G:\\download\_vb\\Assignement\\Basic statistics 1\\Q7.csv")

# For Points

mean(df$Points) ## mean

# 3.596563

median(df$Points) ## median

# 3.695

var(df$Points) ## variance

#0.2858814

sd(df$Points) or sqrt(var(df$Points)) ## standard divation

#0.5346787

range(df$Points) ## max and min val

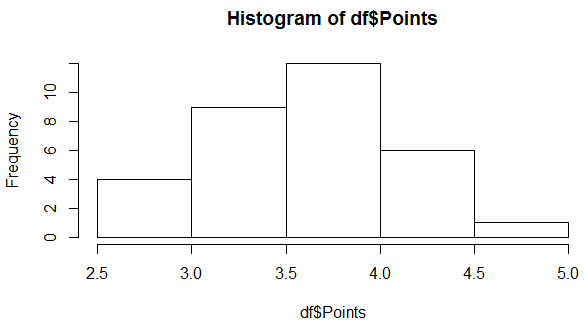
#2.76 4.93

# Visualization

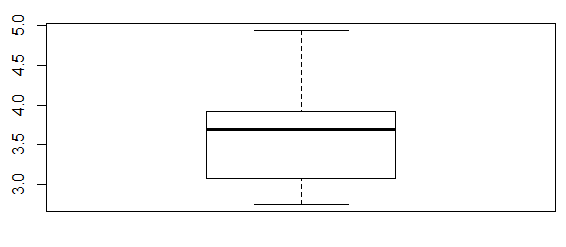
hist(df$Points)

boxplot(df$Points)

Histogram--points



Box Plot- points



df<-read.csv("G:\\download\_vb\\Assignement\\Basic statistics 1\\Q7.csv")

# For Score

mean(df$Score) ## mean

# 3.21725

median(df$Score) ## median

# 3.325

var(df$Score) ## variance

#0.957379

sd(df$Score) or sqrt(var(df$Score)) ## standard divation

#0.9784574

range(df$Score) ## max and min val

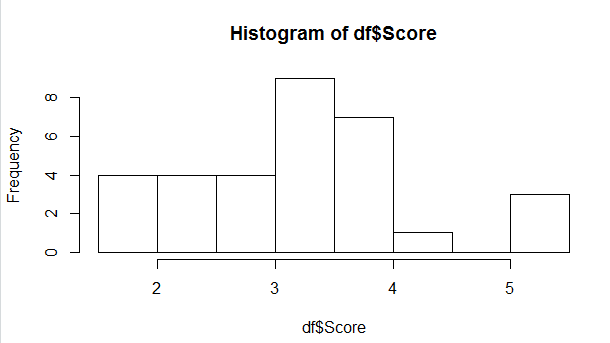
#1.513 5.424

# Visualization

hist(df$Score)

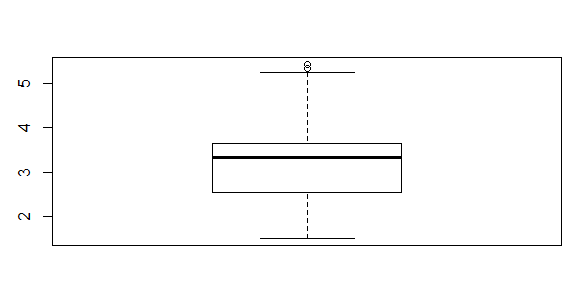
boxplot(df$Score)

Histogram Plot for Score



Box Plot for Score:

Two outliers are present as per graph which can be found by Q3+1.5\*IQR



df<-read.csv("G:\\download\_vb\\Assignement\\Basic statistics 1\\Q7.csv")

# For Weight

mean(df$Weigh) ## mean

# 17.84875

median(df$Weigh) ## median

# 17.71

var(df$Weigh) ## variance

#3.193166

sd(df$Weigh) or sqrt(var(df$Weigh)) ## standard divation

#1.786943

range(df$Weigh) ## max and min val

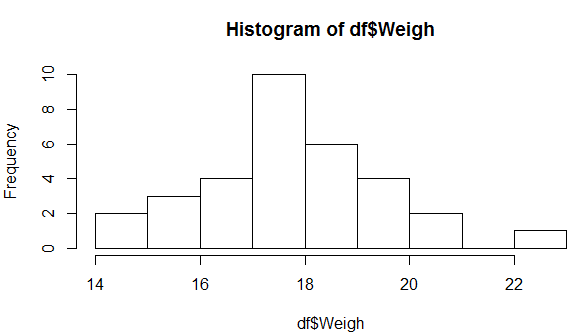
#14.5 22.9

# Vizuliazation

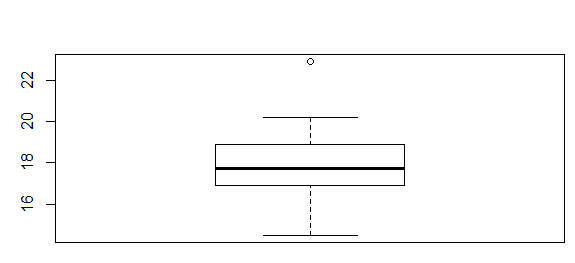
hist(df$Weigh)

boxplot(df$Weigh)

Histogram for weigh



Box Plot for Weigh



Q8) Calculate Expected Value for the problem below

1. The weights (X) of patients at a clinic (in pounds), are

108, 110, 123, 134, 135, 145, 167, 187, 199

Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

Ans.) Based on mean and standard deviation calculation it could be between 112.8103 and 177.8564

Mean Value- 145.3333

Standard deviation- 32.52307

**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data**

**Cars speed and distance**

**Use Q9\_a.csv**

library(e1071)

df<-read.csv("G:\\download\_vb\\Assignement\\Basic statistics 1\\Q9\_a.csv")

skewness(df$speed)

##-0.1105533

plot(density(df$speed))

kurtosis(df$speed)

##-0.6730924

hist(df$speed, # histogram

col="peachpuff", # column color

border="black",

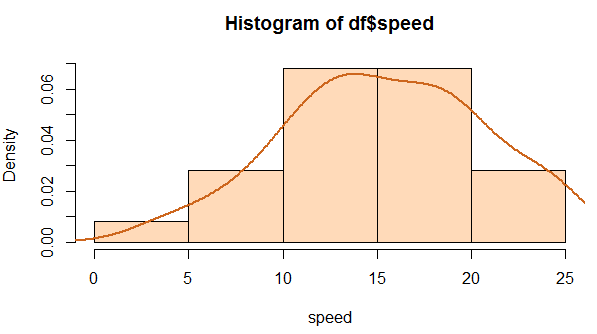
prob = TRUE, # show densities instead of frequencies

xlab = "speed")

lines(density(df$speed), # density plot

lwd = 2, # thickness of line

col = "chocolate3")



library(e1071)

df<-read.csv("G:\\download\_vb\\Assignement\\Basic statistics 1\\Q9\_a.csv")

skewness(df$dist)

## 0.7591268

kurtosis(df$dist)

## 0.1193971

hist(df$dist, # histogram

col="peachpuff", # column color

border="black",

prob = TRUE, # show densities instead of frequencies

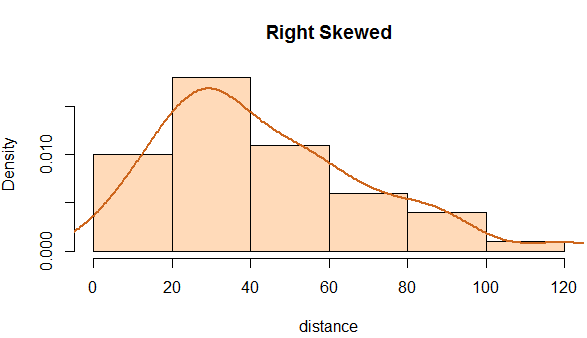
xlab = "distance",

main=”Right Skewed”)

lines(density(df$dist), # density plot

lwd = 2, # thickness of line

col = "chocolate3")



**SP and Weight(WT)**

**Use Q9\_b.csv**

library(e1071)

df<-read.csv("G:\\download\_vb\\Assignement\\Basic statistics 1\\Q9\_b.csv")

skewness(df$SP)

## 1.552258

kurtosis(df$SP)

## 2.583072

hist(df$SP, # histogram

col="peachpuff", # column color

border="black",

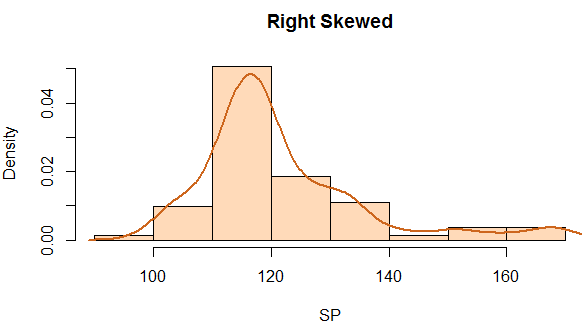
prob = TRUE, # show densities instead of frequencies

xlab = "SP",main="Right Skewed")

lines(density(df$SP), # density plot

lwd = 2, # thickness of line

col = "chocolate3")



library(e1071)

df<-read.csv("G:\\download\_vb\\Assignement\\Basic statistics 1\\Q9\_b.csv")

skewness(df$WT)

## -0.5921721

kurtosis(df$WT)

## 0.7257402

hist(df$WT, # histogram

col="peachpuff", # column color

border="black",

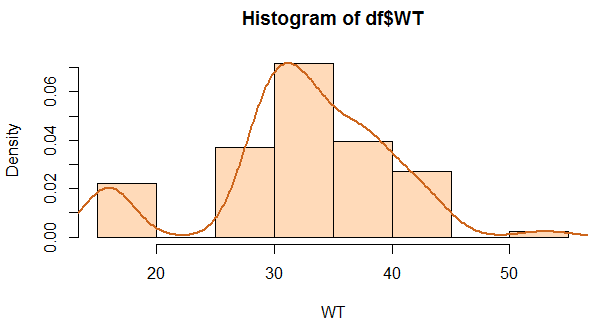
prob = TRUE, # show densities instead of frequencies

xlab = "WT")

lines(density(df$WT), # density plot

lwd = 2, # thickness of line

col = "chocolate3")



**Q10) Draw inferences about the following boxplot & histogram**



**Inferences on above histogram**

Weigh is continous datatype that’s why we are abe to draw histogram

Based on Histogram we have on X axis Weight of Chicks

Y Axis give howmany timesthe value at tha point occures

The data looks right skewed as number of chicks decreasing as weight is increasing

Major data distribution is between 0 to 200 range, though maximum chicks whose weights are beween 50 t0 150 has more numbers in count.

Inter Quartile Range(IQR)=Q3-Q1

Outliers

Q3 + IQR \*1.5

Median (50% Values)

Upper Quartile (75% Values) (Q3)

Lowe Quartile (25% values) (Q1)

Whisker

Q1 - IQR \*1.5



**Q11)** Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%, 98%,96% confidence interval?

**For 94% confidence**

s=30

n=2000

p=0.94 + ((1-0.94)/2)

err=(abs(qt(p,n-1))\*s)/sqrt(n)

err=1.262391

confidence interval will be between **198.7383 and 201.2617**

**For 98% confidence**

s=30

n=2000

p=0.98 + ((1-0.98)/2)

err=abs(qt(p,n-1))\*(s/sqrt(n))

err=1.561814

confidence interval will be between **198.438 and 201.5618**

**For 96% confidence**

s=30

n=2000

p=0.96 + ((1-0.96)/2)

err=abs(qt(p,n-1))\*(s/sqrt(n))

err= 1.378596

confidence interval will be between **198.6214 and 201.378**

**Q12)**Below are the scores obtained by a student in tests

**34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56**

1. Find mean,median,variance,standard deviation.
2. What can we say about the student marks?

ANS) > marks<-c(34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56)

> mean(marks)

[1] 41

> median(marks)

[1] 40.5

> var(marks)

[1] 25.52941

> sd(marks)

[1] 5.052664

1. From this we can say students marks can varies between 35.94 and 46.052

as per mean and deviation for 1 deviation

so major marks will be near 41 means with deviation of 5

Data looks normally distributed as mean and median are having almost similar value

Q13) What is the nature of skewness when mean, median of data are equal?

Data will be normally distributed without skewness

Q14) What is the nature of skewness when mean >median ?

If the mean is greater than the median, the distribution is positively skewed.

Q15) What is the nature of skewness when median > mean?

If the mean is greater than the median, the distribution is negatively skewed.

Q16) What does positive kurtosis value indicates for a data ?

Values are highly distributed at one value this happens when we have similar data at very frequent times. As Result we have pick at middle and at the tails less values are distributed

Q17) What does negative kurtosis value indicates for a data?

When there is flat distribution of data happens. We have that type of data where all values have similar frequency of data.

Q18) Answer the below questions using the below boxplot visualization.



What can we say about the distribution of the data?

ANS.) Major data I distributed between 10 to 16.

75% of values are covered at the value of 18 at upper quartile and 25% of values covered at value of 10 at lower quartile. Median of the data is around 15.5

Data does not look normally distributed.

What is nature of skewness of the data?

ANS.) Data is negatively skewed.

If the mean and median both are same then it will be normally distributed

If mean>median than it will be positive skewed and mean<median than negatively skewed.

What will be the IQR of the data (approximately)?

ANS.) IQR=Q3-Q1

So it will be 8.

Q19) Comment on the below Boxplot visualizations?



Draw an Inference from the distribution of data for Boxplot 1 with respect Boxplot 2.

ANS.) IQR range for plot 2 is higher than plot1

Plot 1 and plot 2 has almost same median value

Both plots have almost similar data distribution in IQR.

From the value we can interpret that 2nd plot refers to big data set than 1st.

Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

Calculate the probability of MPG ofCars for the below cases.

MPG<- Cars$MPG

a.)P(MPG>38)

Ans.) 1-pnorm(38,mean(df$MPG),sd(df$MPG))

**0.3475939**

1. P(MPG<40)

ANS.) pnorm(40,mean(df$MPG),sd(df$MPG))

**0.7293499**

c. P (20<MPG<50)

ANS.) (pnorm(50,mean(df$MPG),sd(df$MPG))) -(1-pnorm(20,mean(df$MPG),sd(df$MPG)))

**0.01311647**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

> mean(df$MPG)

[1] 34.42208

> median(df$MPG)

[1] 35.15273

> skewness(df$MPG)

[1] -0.1714104

> kurtosis(df$MPG)

[1] -0.7054604

> shapiro.test(df$MPG)

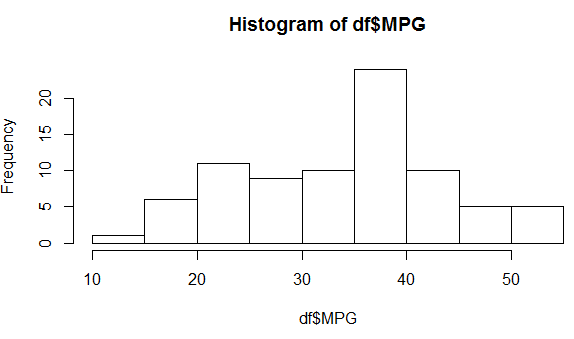
Shapiro-Wilk normality test

data: df$MPG

W = 0.97797, p-value = 0.1764

Here from the data mean and median value are not same but near to each other.

Though Skewness will be there. Skewness and Kurtoziz values are between -2 to 2 as well as p value is greater than 0.05 so it will not reject the null hypothesis. So Data will be some what near to normally distributed as we can see in Histogram also.



1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

> mean(df$Waist)

[1] 91.90183

> median(df$Waist)

[1] 90.8

> skewness(df$Waist)

[1] 0.130389

> kurtosis(df$Waist)

[1] -1.141846

> shapiro.test(df$Waist)

Shapiro-Wilk normality test

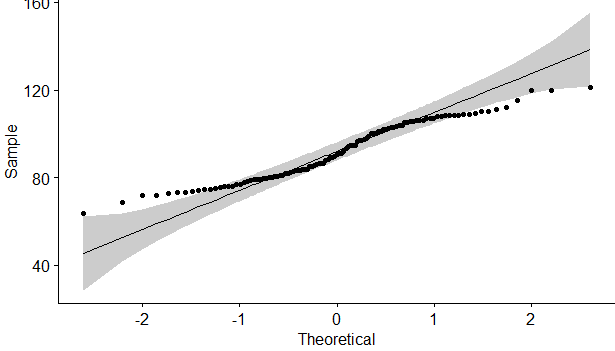
data: df$Waist

W = 0.95586, p-value = 0.00117

Here we can see p-value is less than 0.05 so there as significant data to reject null hypothesis so chances are there that data is not normally distributed.

> library(ggpubr)

> ggqqplot(df$Waist)



Q 22) Calculate the Z scoresof 90% confidence interval,94% confidence interval, 60% confidence interval

90% confidence

> qnorm(0.95)

[1] 1.644854

94% confidence

> qnorm(0.97)

[1] 1.880794

60% confidence

> qnorm(0.80)

[1] 0.8416212

Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

95% confidence

> abs(qt(0.025,24))

[1] 2.063899

96% confidence

> abs(qt(0.02,24))

[1] 2.171545

99% confidence

> abs(qt(0.005,24))

[1] 2.79694

Q 24**)**A Government companyclaims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

Hint:

rcode🡪pt(tscore,df)

df 🡪 degrees of freedom

p-values-🡪 **0.3216725**