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

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

**ANS:-**The Sample Space for above Problem is

S = (HHH,HHT,HTH,THH,HTT,THT,TTH,HHH)

Event set = (HHT,HTH,THH)

Therefore, Probability = **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 2 and 3

**ANS:- Sample space :- {(1,1),(1,2),(1,3),(1,4),(1,5),(1,6)**

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

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

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

**(6,1), (6,2),(6,3),(6,4),(6,5),(6,6)}**

1. Zero
2. Event=(11,12,13,21,22,31)

Therefore , Probability = 6/36

1. Event = (15,24,33,42,51,66)

Therefore, Probability = 6/36

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:-** Total number of balls = (2 + 3 + 2) = 7.  
 n(S)= Number of ways of drawing 2 balls out of 7  = 7C2 `=21  
 Let E = Event of drawing 2 balls, none of which is blue.  
 n(E)= Number of ways of drawing 2 balls out of (2 + 3) balls.= 5C2=10  
 so,

Probability=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

**ANS:-** Expected Value = ∑xP(x) = 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.



**ANS:-** pnt<-read.csv(file.choose())

##Point##

mean(pnt$Points)

median(pnt$Points)

sd(pnt$Points)

var(pnt$Points)

range(pnt$Points)

####Score##

mean(pnt$Score)

median(pnt$Score)

sd(pnt$Score)

var(pnt$Score)

range(pnt$Score)

###Weigh##

mean(pnt$Weigh)

median(pnt$Weigh)

sd(pnt$Weigh)

var(pnt$Weigh)

range(pnt$Weigh)

|  |  |  |  |
| --- | --- | --- | --- |
|  | point | Score | weigh |
| Mean | 3.59 | 3.22 | 17.85 |
| Median | 3.69 | 3.33 | 17.71 |
| variance | 0.29 | 0.96 | 3.19 |
| Standard deviation | 0.53 | 0.98 | 1.79 |
| Range | 2.76 - 4.93 | 1.513 - 5.424 | 14.5 - 22.9 |

**Inference:- We can say Mean is influenced by Extreme values Whereas median is not it is the middle value .**

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:-** Expected Value = ∑X/n =(108+ 110+123+134+ 135+ 145+167+187+199)/9 =1308/9 = 145.33

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

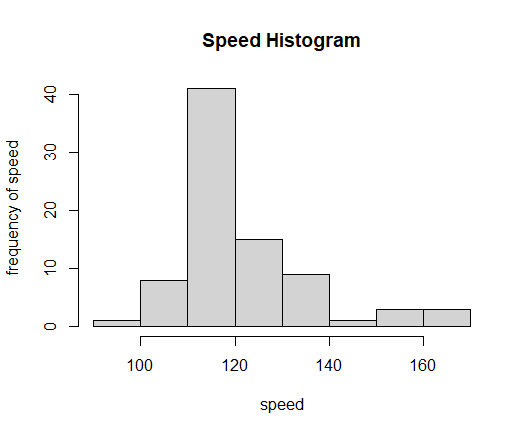
**Cars speed and distance .**

**ANS:-**

****

For speed column

|  |  |
| --- | --- |
| > abc<-read.csv("E:/Data science Excelr/Assigments/BAsic Statistics 1/Q9\_a.csv")  > library(moments)  > skewness(abc$speed)  [1] -0.1139548 ###as the value is negative, it is negatively or left Skewed  > kurtosis(abc$speed)  [1] 2.422853 ##It is playkurtic because it is less than 3   |  | | --- | | > | |
|  |
|  |
| |  | | --- | | > hist(abc$speed, main = "Speed Histogram" ,xlab = "speed",ylab = "frequency of speed") | |  | | E:\Data science Excelr\Assigments\BAsic Statistics 1\Rplot.png  The Skewness value is negative, that indicates the distribution is skewed  towards left that tells Mean of distribution is less than the Median.  Kurtosis Value is less than 3, that tells us that the distribution has broad peak  and thin tails as evident from the histogram which indicates it’s a platykurtic  For distance column   |  | | --- | | > skewness(abc$dist)  [1] 0.7824835 ##Positive/right skewed  > kurtosis(abc$dist)  [1] 3.248019 #  > hist(abc$speed, main = "Distance Histogram" ,xlab = "distance",ylab =  "frequency of Distance") | | E:\Data science Excelr\Assigments\BAsic Statistics 1\dist hist.png  The Skewness value is Positive, that indicates the distribution is skewed towards right that tells Mean of distribution is more than the Median. Kurtosis Value is more than 3, that tells us that the distribution has sharp peak and wide tails as evident from the histogram which says it is a leptokurtic.  **Set_3_@**  For SP  ANS:- > def<-read.csv("E:/Data science Excelr/Assigments/BAsic Statistics 1/Q9\_b.csv")  > skewness(def$SP)  [1] 1.581454  > kurtosis(def$SP)  [1] 5.723521  > hist(def$SP, main = "Speed Histogram" ,xlab = "speed",ylab = "frequency of speed") | |  | |  | |  | | |

****

The Skewness value is Positive, that indicates the distribution is skewed towards right that tells Mean of distribution is more than the Median. Kurtosis Value is more than 3, that tells us that the distribution has sharp peak and wide tails as evident from the histogram which says it is a leptokurtic.

For WT:

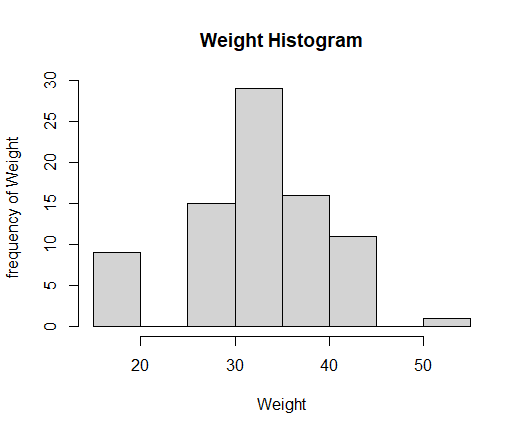
> skewness(def$WT)

[1] -0.6033099

> kurtosis(def$WT)

[1] 3.819466

> hist(def$WT, main = "Weight Histogram" ,xlab = "Weight",ylab = "frequency of Weight")



The Skewness value is negative, that indicates the distribution is skewed

towards left that tells Mean of distribution is less than the Median. Kurtosis Value is more than 3, that tells us that the distribution has sharp peak and wide tails as evident from the histogram which says it is a leptokurtic.

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



**ANS:-**

**The above histogram states that it is skewed towards right i.e here mean>median**.



ANS:-

**The above shown boxplot state that it has many outliers towards upper extreme**

**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 ?

**ANS:-**

**94% Confidence:**

Mean = 200

Sd = 30

n = 2000

CI = Mean ± Z\*Sd/sqrt(n)

=200 ± 1.88\*30/sqrt(2000)

=200 ± 1.262

=198.74 – 201.26

**98% Confidence:**

Mean = 200

Sd = 30

n = 2000

CI = Mean ± Z\*Sd/sqrt(n)

=200 ± 2.33\*30/sqrt (2000)

|  |  |
| --- | --- |
| = | 200 ± 1.561 |

=198.44-201.56

**96% Confidence:**

Mean = 200

Sd = 30

n = 2000

CI = Mean ± Z\*Sd/sqrt(n)

=200 ± 2.05\*30/sqrt (2000)

|  |  |
| --- | --- |
| = | 200 ± 1.378 |

=198.62-201.38

**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:-**

1. Mean = 41

Median = 40.5

Variance = 25.53

Standard Deviation = 5.05

2.Mean > Median, This tells that the distribution is slightly skewed towards right.

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

**ANS:-** Skewness is 0 and it is Symmetric

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

**ANS:-** The nature of skewness is towards right

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

**ANS:-** The nature of skewness is towards left

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

**ANS:-** It has a Sharp peak and has a thick tail.

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

**ANS:-** It has a broad peak and has a thin tail.

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



What can we say about the distribution of the data?

**ANS:-** The data is not a normal distribution is the skewed towards left

What is nature of skewness of the data?

**ANS:-** The nature of skewness is towards left.

What will be the IQR of the data (approximately)?   
**Ans:-** Q3-Q1=18-10=8

Q19) Comment on the below Boxplot visualizations?



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

**ANS:- In above Boxplots we can say the both are normally distributed**

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

Data \_set: Cars.csv

Calculate the probability of MPG of Cars for the below cases.

MPG <- Cars$MPG

* 1. P(MPG>38)

**ANS:-**

> cars<- read.csv("E:/Data science Excelr/Assigments/BAsic Statistics 1/Cars.csv")

> mean(cars$MPG)

[1] 34.42208

> sd(cars$MPG)

[1] 9.131445

> pnorm(38,34.42,9.13)

[1] 0.652513

> 1-0.65

[1] 0.35

So,P(MPG>38)=0.35

* 1. P(MPG<40)

**ANS:-**

> #for(mpg<40)

> pnorm(40,34.42,9.13)

[1] 0.7294571

So, P(MPG<40)= 0.729

* 1. P (20<MPG<50)

**ANS:-**

|  |
| --- |
| > #for(20<mpg<50)  > pnorm(50,34.42,9.13)-pnorm(20,34.42,9.13)  [1] 0.8989178 |

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

ANS:-

cars<- read.csv("E:/Data science Excelr/Assigments/BAsic Statistics 1/Cars.csv")

model.car<-lm(MPG~.,data = cars)

summary(model.car)

#install.packages("carData")

#install.packages("car")

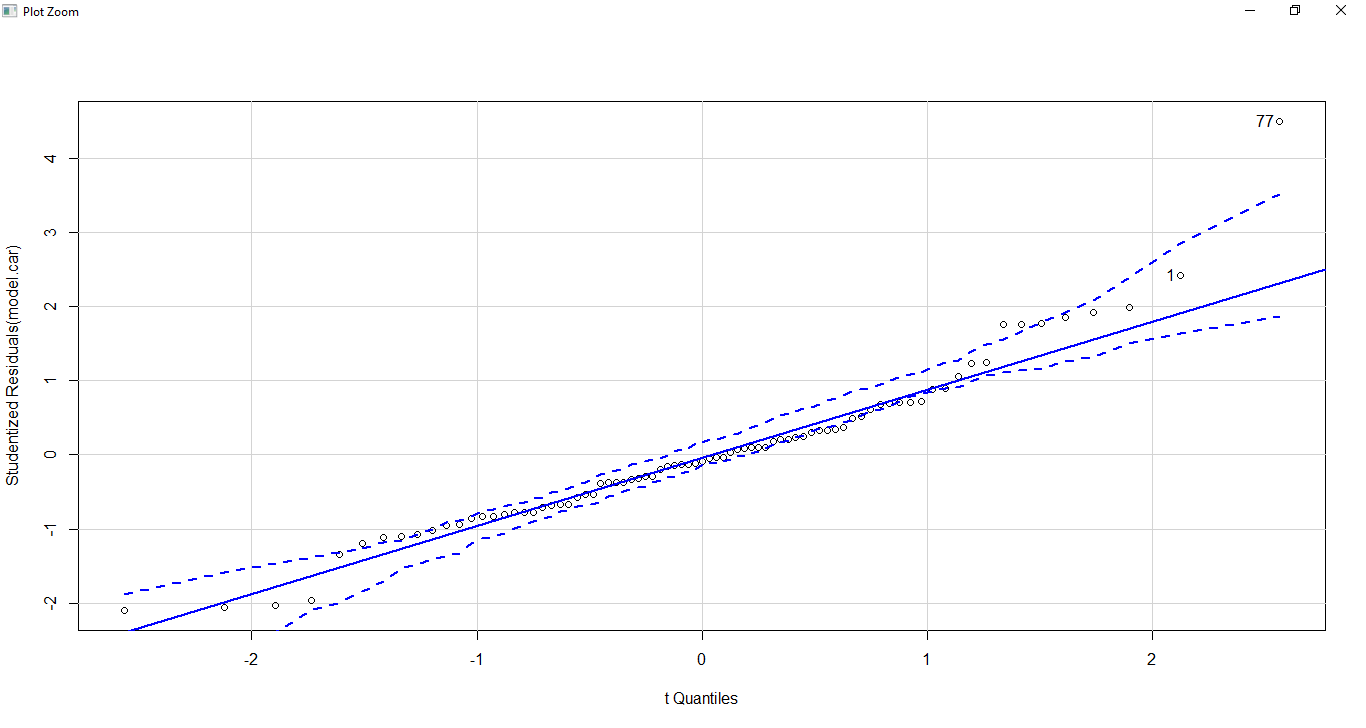
library(carData)

library(car)

car::vif(model.car)

#QQplot

qqPlot(model.car)



**It follows a normal distribution as shown in qqplot**

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

Dataset: wc-at.csv

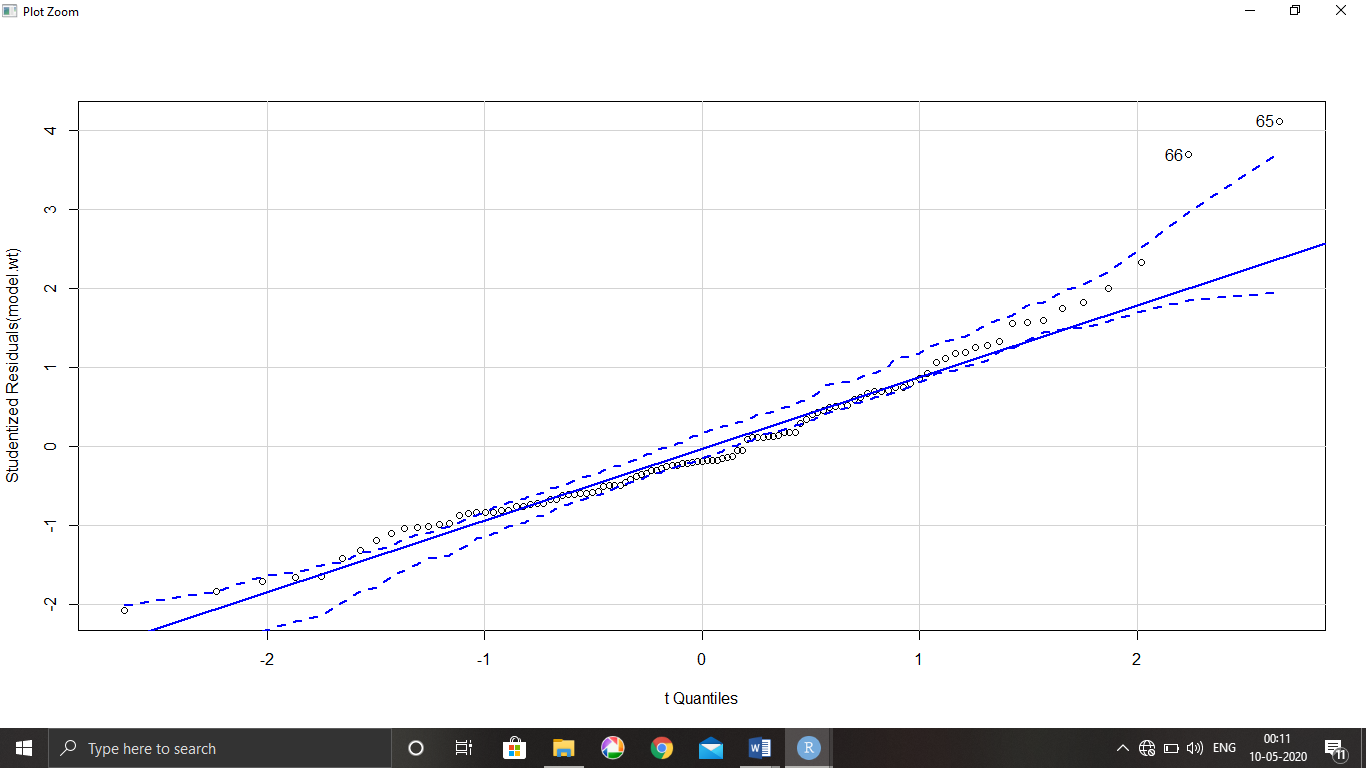
**ANS:- For Waist: It follows a normal distribution**

xyz<-read.csv("E:/Data science Excelr/Assigments/BAsic Statistics 1/wc-at.csv")

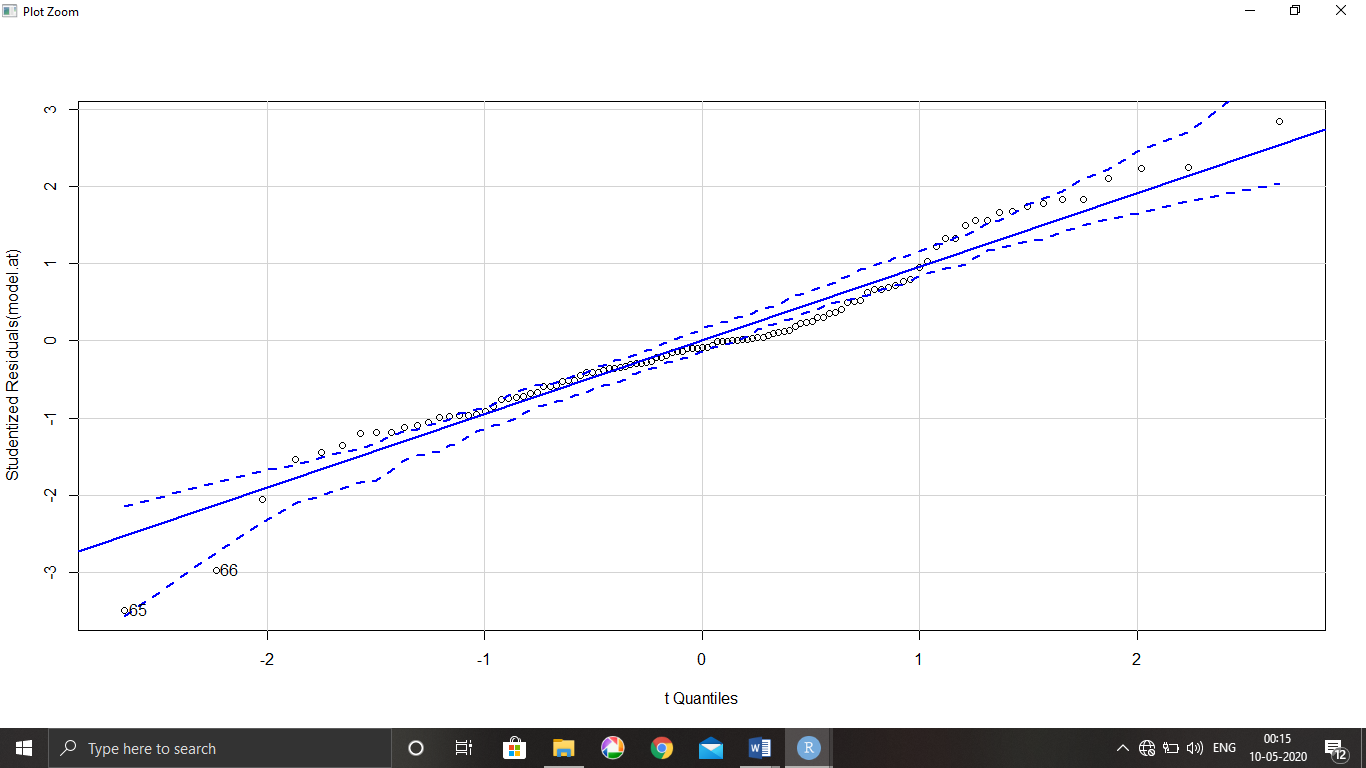
model.wt<-lm(Waist~.,data = xyz)

car::vif(model.wt)

qqPlot(model.wt)



**For AT: It Also follows a normal Distribution**



Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence .

**ANS:-**

|  |
| --- |
| > #For 90%  > qnorm(0.95)  [1] 1.644854  > #for 94%  > qnorm(0.97)  [1] 1.880794  > #For 60%  > 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.

**ANS:-**

> #For 95%

> qt(0.975,24)

[1] 2.063899

> #FOR 96%

> qt(0.98,24)

[1] 2.171545

> #For 99%

> qt(0.995,24)

[1] 2.79694

Q 24**)** A Government company claims 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

**ANS:-**

|  |
| --- |
| > mean<- 260  > mew<- 270  > Std<- 90  > n<-18  > q1<-260-270 #-10  > q2<-Std/sqrt(n) #21.213  > q1/q2 #-0.471  [1] -0.4714045  > pt(-0.471,17)  [1] 0.321814 |

So, probability = 0.321