|  |  |
| --- | --- |
| 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: n(A)={HHT,THH,HTH}=3

n(S)={HHH,HHT,HTT,TTT,THH,TTH,HTH,THT} =8

Required Probability P(A)= **3/8**

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

1. Equal to 1 : Sample Space==

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 11 | 12 | 13 | 14 | 15 | 16 |
| 21 | 22 | 23 | 24 | 25 | 26 |
| 31 | 32 | 33 | 34 | 35 | 36 |
| 41 | 42 | 43 | 44 | 45 | 46 |
| 51 | 52 | 53 | 54 | 55 | 56 |
| 61 | 62 | 63 | 64 | 65 | 66 |

Ans: **Required Probability Zero** because it starts with 1

1. Less than or equal to 4   
   Ans: n(s)=36

n(A)={(1,1)(1,2),(1,3)(3,1)(2,2)(2,1)}

**Required Probability P(A)= 6/36 = 16.66% or 0.1666**

1. Sum is divisible by 2 and 3 means divisible by 6

(1,5)(5,1)(6,6) (3,3)(4,2)(2,4)   
Ans:

**Required Probability P(A)= 6/36 = 16.66% or 0.1666**

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: The number of possible pairs of balls (assuming the ball is drawn and not thrown back in), which is (7C2)=7!/2!\*(7-2)! = 21 .

Now calculate the number of possible pairs given only red and green ball. This gives us (5C2)= 5!/2!\*(5-2)!=10

**the probability is number of red and green ball pairs /number of all possible pairs=10/21 = 0.4761**

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.

**1.** **mean** is the average of the numbers. It is easy to calculate: add up all the numbers, then divide by how many numbers there are.

2.**Median**: to find it Think Middle. To determine the median, the list of numbers should be arranged in order from lowest to highest. ...

**3. Mode**. Mode refers to the number in a list that occurs most often. ...

**4.Variance:-** The average of the squared differences from the Mean.

To calculate the variance follow these steps

i.Work out the Mean (the simple average of the numbers)

ii.Then for each number: subtract the Mean and square the result (the squared difference).

iii.Then work out the average of those squared differences.

**5. Standard Deviation**

The Standard Deviation is a measure of how spread out numbers are.

Its symbol is σ (the greek letter sigma).

The formula is : it is the square root of the Variance

**6. Range=L–S**  
where L – the largest/maximum value attained by the random variable under consideration and S – the smallest/minimum value.

|  |
| --- |
| **> #Points**  > df <- read.csv(file.choose())  > df[,2]  [1] 3.90 3.90 3.85 3.08 3.15 2.76 3.21  [8] 3.69 3.92 3.92 3.92 3.07 3.07 3.07  [15] 2.93 3.00 3.23 4.08 4.93 4.22 3.70  [22] 2.76 3.15 3.73 3.08 4.08 4.43 3.77  [29] 4.22 3.62 3.54 4.11  > ##MEAN  > Points.mean<-mean(df[,2])  > Points.mean  [1] 3.596563  > ##median  > Points.median<-median(df[,2])  > Points.median  [1] 3.695  > ##mode  > mode <- function(x) {  + ux <- unique(x)  + ux[which.max(tabulate(match(x, ux)))]  + }  > Points.mode<-mode(df[,2])  > print(Points.mode)  [1] 3.92  > ##Varience  > var(df[,2])  [1] 0.2858814  > ##Standard Deviation  > Points.Standard\_Deviation<-sd(df[,2])  > Points.Standard\_Deviation  [1] 0.5346787  > ##Range  > Points.Range=(max(df[,2])-min(df[,2]))  > Points.Range  [1] 2.17 |
|  |
| |  | | --- | |  |   > df <- read.csv(file.choose()) |

**> #Score**

> df[,3]

[1] 2.620 2.875 2.320 3.215 3.440 3.460

[7] 3.570 3.190 3.150 3.440 3.440 4.070

[13] 3.730 3.780 5.250 5.424 5.345 2.200

[19] 1.615 1.835 2.465 3.520 3.435 3.840

[25] 3.845 1.935 2.140 1.513 3.170 2.770

[31] 3.570 2.780

> ##MEAN

> Score.mean<-mean(df[,3])

> Score.mean

[1] 3.21725

> ##median

> Score.median<-median(df[,3])

> Score.median

[1] 3.325

> ##mode

> mode <- function(x) {

+ ux <- unique(x)

+ ux[which.max(tabulate(match(x, ux)))]

+ }

> Score.mode<-mode(df[,3])

> print(Score.mode)

[1] 3.44

> ##Varience

> var(df[,3])

[1] 0.957379

> ##Standard Deviation

> Score.Standard\_Deviation<-sd(df[,3])

> Score.Standard\_Deviation

[1] 0.9784574

> ##Range

> Score.Range=(max(df[,3])-min(df[,3]))

> Score.Range

[1] 3.911

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  | | --- | | **> #weigh**  > df <- read.csv(file.choose())  > df[,4]  [1] 16.46 17.02 18.61 19.44 17.02 20.22 15.84 20.00 22.90 18.30 18.90 17.40 17.60 18.00  [15] 17.98 17.82 17.42 19.47 18.52 19.90 20.01 16.87 17.30 15.41 17.05 18.90 16.70 16.90  [29] 14.50 15.50 14.60 18.60  > ##MEAN  > weigh.mean<-mean(df[,4])  > weigh.mean  [1] 17.84875  > ##median  > weigh.median<-median(df[,4])  > weigh.median  [1] 17.71  > ##mode  > mode <- function(x) {  + ux <- unique(x)  + ux[which.max(tabulate(match(x, ux)))]  + }  > weigh.mode<-mode(df[,4])  > print(weigh.mode)  [1] 17.02  > ##Varience  > var(df[,4])  [1] 3.193166  > ##Standard Deviation  > weigh.Standard\_Deviation<-sd(df[,4])  > weigh.Standard\_Deviation  [1] 1.786943  > ##Range  > weigh.Range=(max(df[,4])-min(df[,4]))  > weigh.Range  [1] 8.4 | |  | | |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | | |  |  |  |  | | --- | --- | --- | --- | |  | Points | 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:-   * The mean is useful for identifying trends in the data because we can compare means over a time period to spot trends. The mean is the most common measure of central tendency. * The **median** divides a sample of data in half; it is the middle score. The median is a useful statistic if we think our data have some extreme cases. The median is not impacted by extreme cases, but the mean is. | | |
|  |
| |  | | --- | |  | |

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: X<-c(108, 110, 123, 134, 135, 145, 167, 187, 199)

Expected\_weight<-mean(X)

Expected\_weight = **145.33**

**EV = ∑X/n = 1308/9 =145.33**

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

**Cars speed and distance**

> ##\*\*\*\*\*\*\*\*\*\*\*\*\*\*For Speed\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

> sp=df['speed']

> install.packages("moments")

Error in install.packages : Updating loaded packages

> library(moments)

> skewness(sp)

speed

-0.1139548

> kurtosis(sp)

speed

2.422853

> ##\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*For Distance\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

> dt=df['dist']

> dt

|  |
| --- |
| > ##\*\*\*\*\*\*\*\*\*\*\*\*\*\*For Speed\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*  > sp=df['speed']  > library(moments)  > skewness(sp)  speed -0.1139548  > kurtosis(sp)  speed 2.422853 |
|  |
| |  | | --- | | > | |

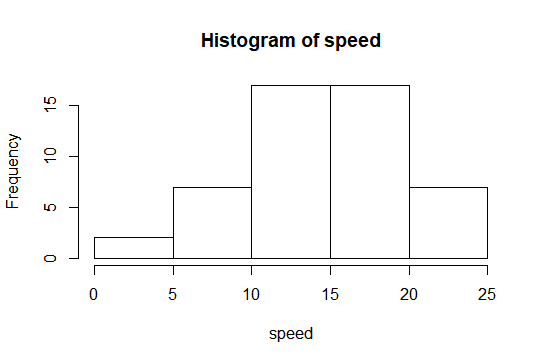
**\*\*\*FOR Speed Column\*\*\*\*\***

**speed -0.1139548** ## left skewness

**speed 2.422853** ## negative Kurtosis

##Draw Histogram

hist(speed)



**Negative Skewness** as Distribution is **skewed towards left.** **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 marked from the histogram and **Kurtosis is negative**

> ##\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*For Distance\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

> dt=df['dist']

> skewness(dt)

dist 0.7824835

> kurtosis(dt)

dist 3.248019

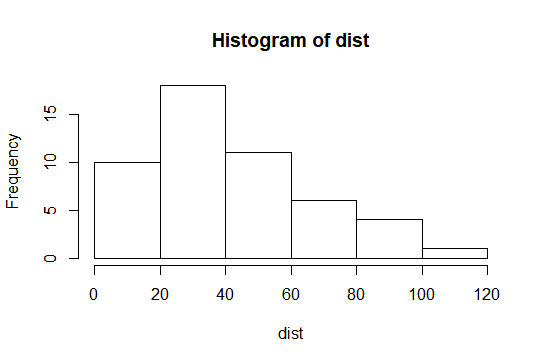
**\*\*\*FOR Distance Column\*\*\*\*\***

**dist 0.7824835** ## Right skewness

**dist 3.248019** ## positive Kurtosis

##Draw Histogram

hist(dist)



**Positive Skewness** as Distribution is **skewed towards Right .** **Mean of distribution is greater than the Median.** **Kurtosis Value is greater than 3**, that tells us that the distribution has sharp peak and wide tails as evident from the histogram .**i.e Positive Kurtosis**

**Q9B Calculate Skewness, Kurtosis & draw inferences on the following data**

**SP and Weight(WT) Use Q9\_b.csv**

> ##\*\*\*\*\*\*\*\*\*\*\*\*\*\*For SP\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

> SP=df['SP']

> library(moments)

> skewness(SP)

SP 1.581454

> kurtosis(SP)

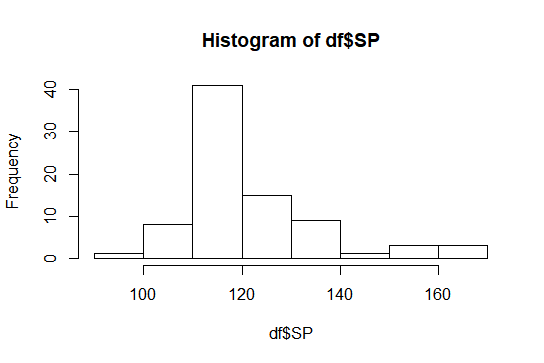
SP 5.723521

**For SP**

**Skewness=**1.**581454 kurtosis=5.723521**

**##Draw Histogram**

**hist(df$SP)**



**Skewness is positive**, that tells us that the distribution is **skewed towards right.** **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.**i.e Positive Kurtosis**

> ##\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*For Weight\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

> Weight=df['WT']

> skewness(Weight)

WT -0.6033099

> kurtosis(Weight)

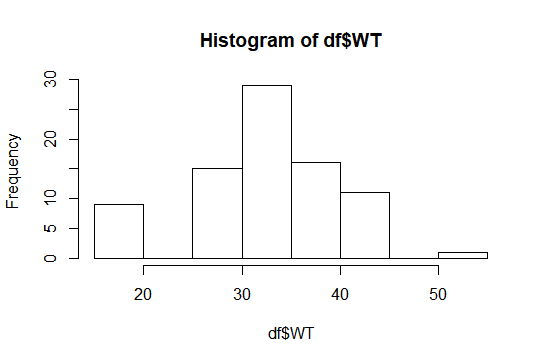
WT 3.819466

**For WT(weight)**

**Skewness= -0.6033099** **kurtosis=3.819466**

**##Draw Histogram**

**hist(df$WT)**



**Skewness is negative**, that tells us that the **distribution is skewed towards left**. **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. **i.e Positive Kurtosis**

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



**Ans: The Distribution is Right Skewed. Mean > Median.**



**Ans:** **The above boxplot suggests that the distribution has lots of outliers towards upper extreme. 7 outliers in above boxplot**

**What is an outlier?**

An outlier is a data point in a data set that is distant from all other observations. A data point that lies outside the overall distribution of the dataset.

**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:**

X-bar = 200

Sd = 30

n = 2000

Interval Estimate = X-bar ± Z\*Sd/sqrt(n)

=200 ± 1.88\*30/sqrt(2000)

=**198.74 – 201.26**

>**98% Confidence:**

X-bar = 200

Sd = 30

n = 2000

Interval Estimate = X-bar ± Z\*Sd/sqrt(n)

=200 ± 2.33\*30/sqrt (2000)

=**198.44-201.56**

>**96% Confidence:**

X-bar = 200

Sd = 30

n = 2000

Interval Estimate = X-bar ± Z\*Sd/sqrt(n)

=200 ± 2.05\*30/sqrt (2000)

=**198.62-201.38**

**For 94% = 200±1.26**

**For 98% = 200±1.56**

**For 96% = 200±1.37**

**Solution :-**

> a <- 200 #sample mean

> s <- 30 #sd

> n <- 2000 # sample size

> #conf interval =94%

> (1-0.94)/2

[1] 0.03

> error <- qnorm(0.03)\*s/sqrt(n)

> error

[1] -1.261675

> left <- a-error

> right <- a+error

> left

[1] 201.2617

> right

[1] 198.7383

> #for 98%

> error <- qnorm(0.01)\*s/sqrt(n)

> error

[1] -1.560562

> left <- a-error

> right <- a+error

> left

[1] 201.5606

> right

[1] 198.4394

> #for 96%

> (1-0.96)/2

[1] 0.02

> error <- qnorm(0.02)\*s/sqrt(n)

> error

[1] -1.377697

> left <- a-error

> right <- a+error

> left

[1] 201.3777

> right

[1] 198.6223

**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 : Mean=41 Median=40.5 Variance=25.52 Standard deviation=5.05**

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

>

> result.mean<-mean(x)

> result.mean #MEAN

[1] 41

> result.median<-median(x)

> result.median #MEDIAN

[1] 40.5

>

> result.variance<-var(x)

> result.variance #VARIANCE

[1] 25.52941

> result.sd<- sd(x)

> result.sd #STANDARD DEVIATION

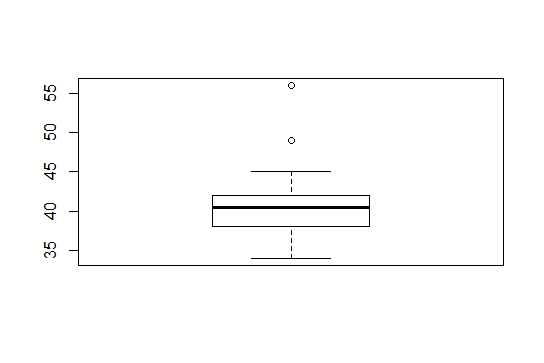
[1] 5.052664

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

**# Draw Boxplot**

**boxplot(x)**

**Two outliers are present.**



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

**Ans: It is called zero skewed.Skewness=0,Symmetric**

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

**Ans: Right skewed distribution**

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

**Ans: Left skewed distribution**

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

**Ans: A distribution with a positive kurtosis value indicates that the distribution has thick tails and a sharper peak than the normal distribution**

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

Ans: **A distribution with a negative kurtosis value indicates that the distribution has lighter(Thin) tails and a flatter (Broad)peak than the normal distribution**

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



What can we say about the distribution of the data?

**Ans: asymmetrical distribution ,Not normal distribution**

What is nature of skewness of the data?

**Ans: Left skewed**

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

**Ans : IQR = 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: **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)
  2. P(MPG<40)

c. P (20<MPG<50)

Ans :

|  |
| --- |
| > names(df)  [1] "HP" "MPG" "VOL" "SP" "WT"  > mpg<-df[,2]  > ##p(mpg>38)  > pnorm(38, mean(mpg), sd(mpg), lower.tail=FALSE)  [1] 0.3475939  > ## p(mpg<40)  > pnorm(40, mean(mpg), sd(mpg))  [1] 0.7293499  > ##P(20<MPG<50)  > pnorm(20, mean(mpg), sd(mpg))-pnorm(50, mean(mpg), sd(mpg))  [1] -0.8988689 |
|  |
| |  | | --- | |  | |

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

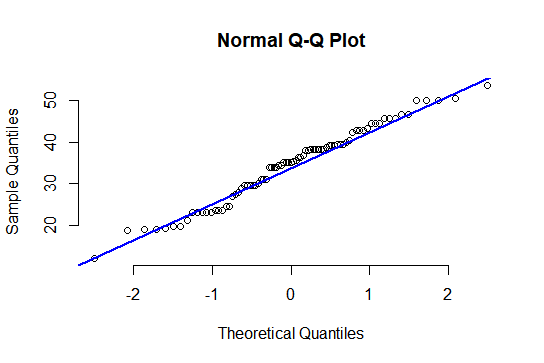
Dataset: Cars.csv

ANS:

names(df)

qqnorm(df$MPG, pch = 1, frame = FALSE)

qqline(df$MPG, col = "Blue", lwd = 2).



Ans: **Follows Normal Distribution as indicated by the qqplot:-** . QQ plot (or quantile- quantile plot) draws the correlation between a given sample and the normal distribution. A 45-degree reference line is also plotted. As all the points fall approximately along this reference line, we can assume normality

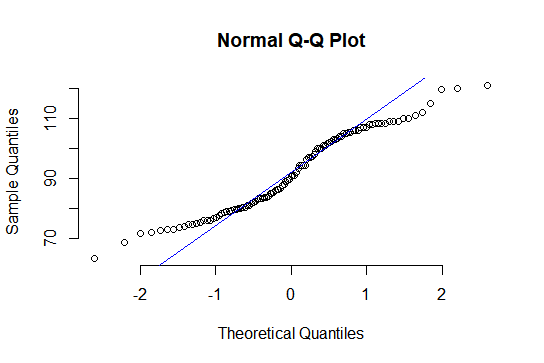
b. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

##Q-Q Plot of waist

qqnorm(wc$Waist, pch = 1, frame = FALSE)

qqline(wc$Waist, col = "Blue") hist(wc$Waist)



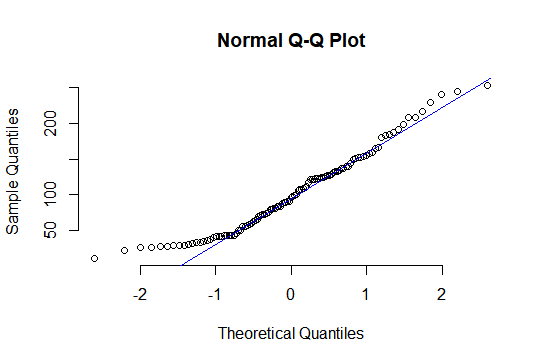
.

**Wc-at $waist follows Normal Distribution :-** As all the points fall approximately along this reference line

##Q-Q Plot of AT

qqnorm(wc$AT, pch = 1, frame = FALSE)

qqline(wc$AT, col = "Blue")



**ANS: Wc-at$AT follows Normal Distribution :-** As all the points fall approximately along this reference line.

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

**Ans: At 90% z score is 1.648, at 94% z = 1.88, at 60% z = 0.841**

> # Z scores

> # for 90%

> p= 0.90+ 0.05

> qnorm(0.95)

[1] 1.644854

> # for 94%

> p= 0.94+0.03

> qnorm(0.97)

[1] 1.880794

> #for 60%

> p= 0.60+0.2

> qnorm(0.8)

[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: At 95% T score = 2.063, at 96% t score is 2.171, at 99% t score = 2.796**

> # T scores --> sample size =25

> df=25-1

[1] 24

> # for 95%

> p=(1-0.95)/2 + 0.95

> qt(0.975,24)

[1] 2.063899

> # for 96%

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

> qt(0.98,24)

[1] 2.171545

> #for 99%

> p=(1-0.99)/2 + 0.99

> 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 =18-1)

Ans:-

> x=260 #Sample mean

> μ= 270 #Population mean

> s= 90 #Sample standard deviation

> n=18

> df= 18-1 = 17 #Degrees of freedom

> t = [ x - μ ] / [ s / sqrt( n ) ]

> t = ( 270 - 260 ) / ( 90 / sqrt( 18) )

> t= -(10)/21.2132 # t\_score

> -(10)/21.2132

[1] -0.4714046

> pt(-0.4714045,17)

[1] 0.3216725

probability that 18 randomly selected bulbs would have an average life of is less thant 260 days =0.3216 **i.e 32.16%**