|  |  |
| --- | --- |
| 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 | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Nominal |
| 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 | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Ratio |
| Years of Education | Ratio |

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

Ans: Total outcomes = 8

2 H, 1 T combinations: HHT, HTH, THH

Probability for 2H & 1T = 3/8.

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

Ans:

1. Equal to 1 = 0/36 = 0
2. Less than or equal to 4 :

Combinations: (1,1), (1,2) , (1,3), (2,1),(2,2),(3,1)

Probability = 6/36 = 1/6.

1. Sum is divisible by 2 and 3

Valid sums are : 6, 12

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

Probability : 6/36 = 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: Total combinations = 7C2 = 7!/(5!\*2!) = 7\*6\*5!/ 5!\*2 = 21

No.of ways we can pick 2 balls without blue :

1. 2 red balls
2. 2 green balls
3. 1 red and 1 green ball

ways to pick 2 red balls out of 2 red balls= 2C2 = 2!/0!\*2! = 1

ways to pick 2 green balls out of 3 green balls = 3C2 = 3

ways to pick 1 red ball out of 2 red balls and 1 green ball out of 3 green balls : = 2C1 \*3C1 = 2 \* 3 = 6

Total no.of ways to pick 2 balls without blue balls randomly : 1+3+6 = 10

Probability to pick 2 balls without blue balls randomly := 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 number of candies for a random child = Expected value = ∑x.p(x)

x p(x) x.p(x)

|  |  |  |  |
| --- | --- | --- | --- |
| 1 | 0.015 | = | 0.015 |
| 4 | 0.2 | = | 0.8 |
| 3 | 0.65 | = | 1.95 |
| 5 | 0.005 | = | 0.025 |
| 6 | 0.01 | = | 0.06 |
| 2 | 0.12 | = | 0.24 |
|  |  | Sum | 3.09 |

Ans: 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:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | points | score | | weigh | | |
| mean | 3.596563 | 3.21725 | | 17.84875 | | |
| median | 3.695 | 3.325 | | 17.71 | | |
| mode | 3.07 & 3.92 | 3.44 | | 17.02 & 18.9 | | |
| variance | 0.276948 | 0.927461 | | 3.09338 | | |
| SD | 0.526258 | 0.963048 | | 1.758801 | | |
| range | (2.76 – 4.93), 2.17 | (1.513 – 5.424) ,3.911 | | (14.5 – 22.9), 8.4 | | |
|  | | |  | |  |  | |

Inference:

Points : The average value of distribution of variable points is 3.596563 and the middle most value is 3.695. mean is less than median so data is slightly skewed negatively. Most occurring value is 3.07 & 3.92, so it is bimodal distribution.

It has a variance of 0.276948 and standard deviation of 0.526258 from the mean value. Data points are ranges between 2.76 – 4.93.

Score: The average value of distribution of variable points is 3.21725 and the middle most value is 3.695. mean is less than median so data is slightly skewed negatively. Most occurring value is 3.44, so it is unimode distribution.

It has a variance of 0.927461and standard deviation of 0.963048 from the mean value. Data points are ranges between 1.513 – 5.424.

Weigh: The average value of distribution of variable points is 17.84875 and the middle most value is 17.71. mean is greater than median so data is slightly skewed positively. Most occurring value is 17.02 & 18.9, so it is bimodal distribution.

It has a variance of 3.09338 and standard deviation of 1.758801 from the mean value. Data points are ranges between 14.5 – 22.9.

**Use Q7.csv file**

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: Since all the weights in the given data sets are unique, the probability of their occurrence is also equal. Hence the probability of each weight to occur is 1/9. Applying the formula to find expected value on these data:

Expected value = ∑x.p(x)

|  |  |  |
| --- | --- | --- |
| Weight | probability | E.V |
| 108 | 0.11111111 | 12 |
| 110 | 0.11111111 | 12.22222 |
| 123 | 0.11111111 | 13.66667 |
| 134 | 0.11111111 | 14.88889 |
| 135 | 0.11111111 | 15 |
| 145 | 0.11111111 | 16.11111 |
| 167 | 0.11111111 | 18.55556 |
| 187 | 0.11111111 | 20.77778 |
| 199 | 0.11111111 | 22.11111 |
|  |  |  |
|  | sum = | 145.3333 |

So, Expected value for weight of any random patient = 145.3333

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

**Cars speed and distance**

**Use Q9\_a.csv**

Answer:

|  |  |  |
| --- | --- | --- |
| Moments | Speed | distance |
| Skewness | -0.1139548 | 0.7824835 |
| Kurtosis | 2.422853 | 3.248019 |

Inferences:

Skewness of Speed is negative; hence data is not normally distributed, the mass is on the right side of the bell curve. Also, median will be greater than the mean of this distribution.

Skewness of distance is positive; hence data is not normally distributed, the mass is on the left side of the bell curve. Also, mean will be greater than the median of this distribution.

Kurtosis is positive for both speed and distance. It indicates the distribution is peaked and possess thick tails. In this distribution more of the values are located in the tails of the distribution rather than around the mean. This type of distribution is called as leptokurtic distributions.

**SP and Weight(WT)**

**Use Q9\_b.csv**

|  |  |  |
| --- | --- | --- |
| Moments | SP | WT |
| Skewness | 1.581454 | -0.6033099 |
| Kurtosis | 5.723521 | 3.819466 |

Inferences:

Skewness of SP is positive; hence data is not normally distributed, the mass is on the left side of the bell curve. Also, mean will be greater than the median of this distribution.

Skewness of WT is negative; hence data is not normally distributed, the mass is on the right side of the bell curve. Also, median will be greater than the mean of this distribution.

Kurtosis is positive for both SP and WT. It indicates the distribution is peaked and possess thick tails. In this distribution more of the values are located in the tails of the distribution rather than around the mean. This type of distribution is called as leptokurtic distributions.

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



Ans : Histogram is plotting the frequency distribution of chickweight$weight. The data is observed between values of 0 to 400 for chickweight$weight and number of occurrences/ frequency is between 0 to 200.

According to this histogram, data is distributed widely between 0 to 200. Most of the chickweight$weight lies between 50 to 100.

The data is positively skewed.



Ans: There are some outliers in the data. 50% of the data lies near to the lower whisker. Data is positively skewed.

**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:

Confidence Interval = t (1-α, n-1) \*s/

n=2000, s= 30, sample mean = 200

s/ = 30/ = 0.6708

t (1-α, n-1) -> t(94%,1999)? t(98%,1999)? t(96%,1999)?

Confidence Interval for 94% confidence level:

qt(0.97,1999) =1.88

2001.88\*0.6708 =200 1.261104

Confidence Interval at 94 % = [198.74 201.26 ]

Confidence Interval for 96% confidence level:

qt(0.98,1999) = 2.05509

2002.05509\*0.6708 = 200 1.379

Confidence Interval at 96 % = [198.62 201.38]

Confidence Interval for 98% confidence level:

qt(0.99,1999) = 2.328

2002.328\*0.6708 = 200 1.5616

Confidence Interval at 98 % = [198.4384 201.5616]

**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.

Ans:

|  |  |
| --- | --- |
| Mean | 41 |
| median | 40.5 |
| variance | 24.11111 |
| Standard deviation | 4.910307 |

1. What can we say about the student marks?

Ans:

Student’s average mark is 41.

50% of the student’s mark is less than or equal to 40.5

Student is not consistent with his scores as there is a significant variance of 24.111 between different subjects.

There is a significant standard deviation of 4.910307 between marks.

The marks ranges between 34 to 56.

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

Ans: When mean and median of data are equal, the distribution becomes symmetric and hence normal distribution. Since it is a symmetric distribution , skewness will be zero.

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

Ans: When mean is greater than median, distribution is positively skewed. Majority of the data lies on the left side of the bell curve.

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

Ans: When median is greater than mean, distribution is negatively skewed. Majority of the data lies on the right side of the bell curve.

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

Ans: It indicates the distribution is peaked and possess thick tails. In this distribution more of the values are located in the tails of the distribution rather than around the mean. This type of distribution is called as leptokurtic distributions.

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

Ans: It indicates that the distribution has lighter tails and flatter peaks than of normal distribution. If the data is normally distributed then kurtosis will be zero. When it deviates the distribution , kurtosis value changes.

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



What can we say about the distribution of the data?

Ans: This is an asymmetric distribution. Median of the distribution is 15 almost. The data is skewed negatively. Most of the data (50%) lies between 10-18 values. 25% of data lies below 10. Since whiskers are not shown here, we cannot confirm about minimum and maximum value also about outliers.

25th percentile is 10 and 75th percentile is 18.

What is nature of skewness of the data?

Ans: The data is negatively skewed. The mass of distribution is towards right side. Median is greater than mean.

What will be the IQR of the data (approximately)?   
  
IQR = Q3-Q1

Ans : 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 the data set shows a normal distribution of data across, with mean and median are equal to 262.5. Hence mean and median are equal there is no distribution and hence normal distribution. Even though the mean and median equal, the number of observations in both data are different. Boxplot1 has got few data which varies between 237.5 (minimum- lower whisker ) to 287.5 (maximum -upper whisker) almost. Boxplot 2 data got values between 200 (minimum)and 325 (maximum). (All figures approximate calculation). 50 % of data 1 lies between 250 to 275 and IQR is 25. IQR of data 2 is 312.5 – 225 = 87.5.

Both data set do not contain any outliers. We could assume boxplot 1 as a sample of boxplot 2 data.

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:

Mean (MPG) = 34.42208

Standard Deviation (MPG) = 9.131445

1. P(MPG>38)

Z= 38-34.42208/9.131445 = 0. 3918240760361586

P(0. 3918240760361586) from z table = 0.65173

P(MPG>38) = 1- 0.65173 = 0.34827 = 34.827%

1. P(MPG<40)

Z= 40-34.42208/9.131445 = 0.6108474617106055

P(0.6108474617106055) = 0.72907

P(MPG<40) = 0.72907 = 72.907%

1. P (20<MPG<50)

Ans:

P(MPG<50) ? P(MPG<20) ?

P(MPG<50)

Z= 50 - 34.42208/9.131445 =1.70596439008284

P(1.70596439008284) = 0.95543

P(MPG<50) = 0.95543

P(MPG<20)

Z= 20- 34.42208/9.131445 = -1.579386395033864

P(-1.579386395033864) = 0. 05821

P(MPG<20) = 0. 05821

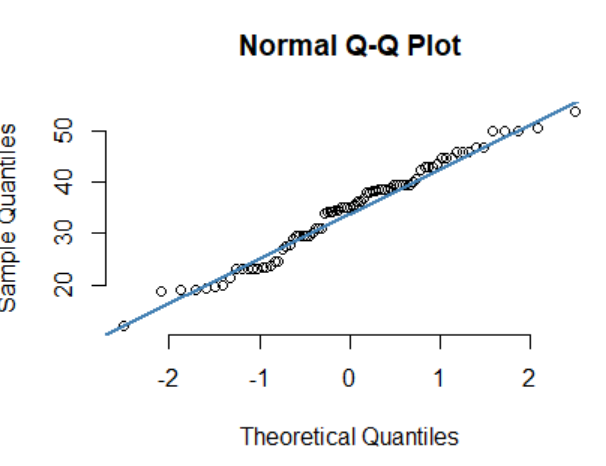
P (20<MPG<50) = P(MPG<50) - P(MPG<20) = 0.95543 - 0. 05821

P (20<MPG<50) = 0.89722

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

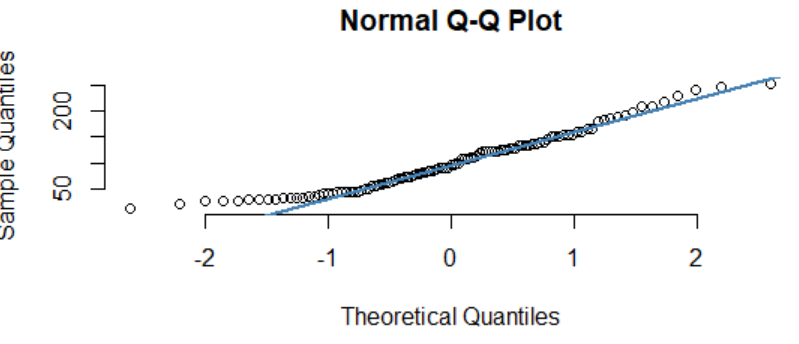


According to the QQplot, the data is normally distributed since there is a straight line is shown.

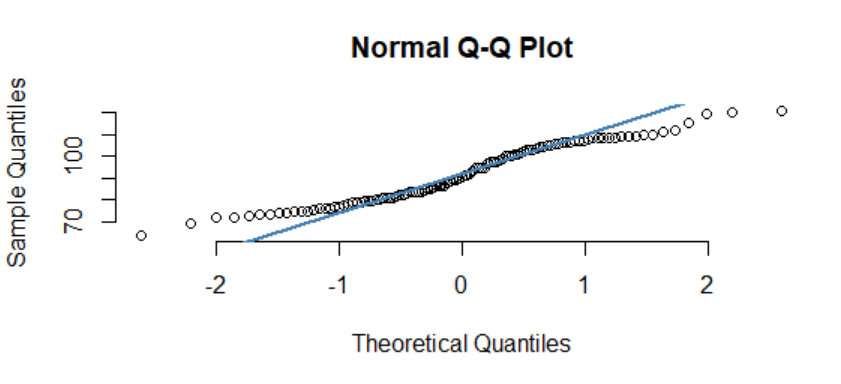
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv

Ans:



The above qqplot shows the normality of AT of wc-at dataset. The data distribution shows a straight line in plot which means it is normally distributed.



The above one shows the normality of Waist data of wc-at dataset. It has also given a straight line but slight deviation at the extremes of data. It is almost normally distributed dataset.

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

Ans:

Considering mean as 0 and standard deviation as 1 in a normal distribution.

Z score of 90% = Z(.95)

qnorm(0.95) = 1.644854

Z score of 90% = 1.644854

Zscore (94%) = qnorm(.970) = 1.880794

Zscore (60%) = qnorm(0.80) = 0.8416212

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

Ans:

t95% = t (0.975,24) =qt(0.975,24) = 2.063899

t96% = t (0.98,24) =qt(0.98,24) = 2.171545

t99% = t (0.995,24) =qt(0.995,24) = 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:

Population mean ,µ = 270

Sample mean , = 260

Standard deviation of sample , s = 90

Number o samples , n= 18

P() ?

Df = n-1 = 17.

tscore =  = 260 – 270/(90) = -0.4714

Probability p, = pt(tscore, df) = pt(-0.4714,17) = 0.3216741

P() = 0.3216741