|  |  |
| --- | --- |
| **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 | Categorical |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Categorical |

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 | Order |
| 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 | Ratio |
| SAT Scores | Interval |
| Years of Education | Interval |

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

Total No. of Events = {HHH, HHT, HTT, TTT, THT, THH, HTH, TTH} = 8

No. Of Expected Events = {HHT, THH, HTH}

Ans:- Probability = 3/8 = 0.375

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

Total No. Of Events = {(1,1)(2,2)(3,3)(4,4)(5,5)(6,6)((1,2)(1,3)(1,4)(1,5)(1,6)

(2,1), (2,3), (2,4), (2,5), (2,6), (3,1), (3,2), (3,4), (3,5), (3,6), (4,1), (4,2), (4,3), (4,5), (4,6), (5,1), (5,2), (5,3), (5,4), (5,6), (6,1), (6,2), (6,3), (6,4), (6,5)}= 36

1. Equal to 1

Ans: - p= 0/36 = 0

1. Less than or equal to 4

Ans: -

Interested Event = (11, 12, 13, 21, 22, 31)

p = 6 /36 = 0.167

The probability that the sum is 6

1. Sum is divisible by 2 and 3

Ans: -

Interested Event = (15, 24, 33, 42, 51, 66)

p = 6/36 =0.167

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 No. of balls = 7

2 balls can be selected in 7C2 ways

If none of the balls are removed then no. of balls left = 5

Now 2 balls can be selected in 5C2 ways

P = 5C2/7C2 = 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 = **∑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=:-

**For Point**

Mean = 3.59

Median = 3.695

Mode = 3.92

Variance = 0.286

Standard Deviation = 0.5346

Range = 2.17

**b.) For Score**

Mean = 3.217

Median = 3.325

Mode= 3.44

Variance = 0.957

Standard Deviation= 0.978

Range = 3.911

**C.) For Weigh**

Mean =17.84

Median = 17.71

Mode =17.02

Variance=3.19

Standard Deviation=1.7869

Range=8.4

Inference Drawn:

* The mean is useful for spotting 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:-Expected Value = **EV = ∑X/n=**  (108+110+123+134+135+145+167+187+199)/9

= 145.33

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

**Cars speed and distance**



**Ans:-**

**For Speed**

Skewness = - 0.8448909

Kurtosis = 2.991396

The “speed” data set is negatively skewed which means that most of the data points lie on the right side of the curve and tail is extended towards Left side. Moreover it is having a slightly high peak.

Or

Skewness is negative, that tells us that the 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 evident from the histogram.

**For Distance**

Skewness = 1.217917

Kurtosis = 4.816933

The “distance” data set is positively skewed which means that most of the data points lie on the left side of the curve and tail is extended towards right side. Moreover it is having high peak.

Or

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.



**For Speed**

Skewness = -0.4076944

Kurtosis = 2.086738

The “speed” dataset is representing negatively skewed distribution, maximum data points lie on the right side of the curve and the tail is extended towards left side as very few data points are lying in the left side. As kurtosis value is positive, which shows that the curve is having slightly higher peak.

Or

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.

**For Weight**

Skewness= -1.287316

Kurtosis= 3.818755

The “weight” dataset is representing negatively skewed distribution, maximum data points lie on the right side of the curve and the tail is extended towards left side as very few data points are lying in the left side. As kurtosis value is positive, which shows that the curve is having much higher peak. If we analyze the histogram and box plot of the dataset, we can find some outlies in the upper extreme as well as lower extreme.

Or

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.

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



Ans:- From the above histogram, it can be said that most of the data points are concentrated on the left side and the tail is in the right side. It implies that it is having +ve ly skewed distribution and there can be some outliers in the datasets.

The Distribution is Right Skewed. Mean > Median.



Ans: - From the above box plot, it can be said that the median is closer to the bottom of the box, whisker is shorter at the lower end of the box, so the distribution is positively skewed. There are few outliers which lies outside the higher extreme value.

The above boxplot suggests that the distribution has lots of 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:- Considering it as a normal distribution,

Given,

n=2000,

= 200 pounds

s= 30

Since population standard deviation is not known,

CI = ± t(1-α,n-1) \*s/n^(1/2)

For 94%,

t(0.97, 1999) = 1.881861

CI = 200 ± 1.881861 \* 30/2000^(1/2) = 200 ± 1.262 = 198.74 to 201.26

For 98%,

t(0.99,1999) = 2.328215

CI = 200 ± 2.328215 \* 30/2000^(1/2) = 200 ± 1.5618 = 198.43 to 201.56

For 96%,

t(0.98,1999) = 2.05509

CI = 200 ± 2.05509 \* 30/2000^(1/2) = 200 ± 1.378596 = 198.62 to 201.388

**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=25.529

Standard deviation = 5.053

1. What can we say about the student marks?

Most of the data points lies around the mean but there are some extreme values (outliers) which is much higher than the mean.

Mean > Median, This implies that the distribution is slightly skewed towards right. No outliers are present.

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

Ans: - Skewness is 0. Symmetric

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

Ans: - Skewness is positive, Right Skewed

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

Ans: - Skewness is negative, Left Skewed

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

Ans: - It indicates that the curve is taller than the curve of normal distribution. Tail part does not have much data. (Sharp Peak, Thick Tails)

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

Ans:- It indicates that the curve is flatter and most of the data point lies on the tail portion of the curve. (Broad Peak, Thin Tails)

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



What can we say about the distribution of the data?

Ans: - The most of the data points are distributed towards right side of the curve.

What is nature of skewness of the data?

Ans: - negatively 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: - The data is spread across wider range in 2 when compared to 1. Inter Quartile Range of 1 is less than 2. Both are Normally Distributed

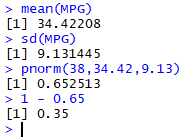
Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

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

MPG <- Cars$MPG

* 1. P(MPG>38)



Ans:- p= 33/81 = 0.407

* 1. P(MPG<40)



Ans:- p = 61/81 = 0.753

c. P (20<MPG<50)



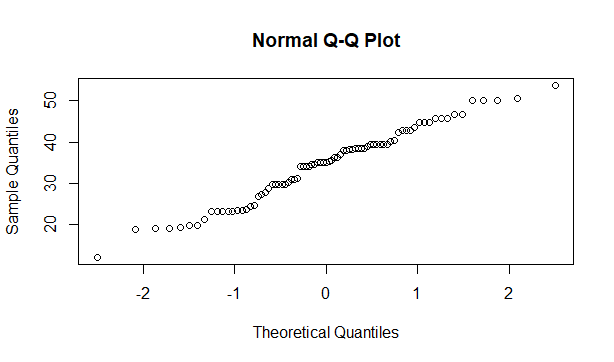
Ans:- p = 12/81 = 4/27 = 0.148

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans:- No it does not follow normal distribution. As from the histogram it is clearly mentioned that the curve is not symmetrical and not bell shaped. Rather it is unsymmetrical and flatter in shape

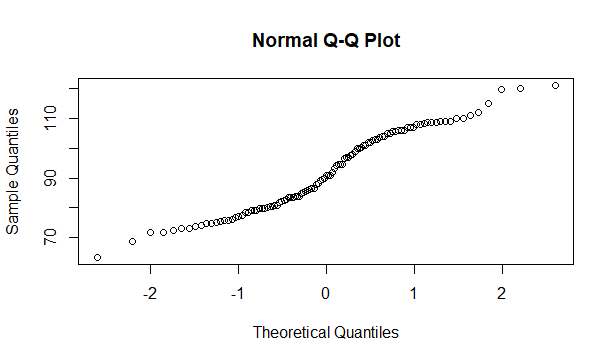


1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

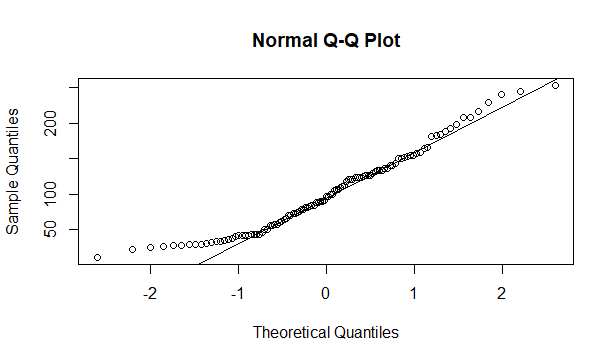
Dataset: wc-at.csv

Ans: -

Waist Circumference datasets does not follow normal distribution as the curve display in the histogram is not symmetrical and bell shaped. It is slight positively skewed data.



Adipose Tissue dataset also does not follow normal distribution as it can be clearly found in histogram and boxplot. Data is positively skewed data.



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

Ans:-

For 90%,

From R,

qnorm(0.95)= 1.644854 = 1.64

For 94%,

qnorm(0.97)= 1.880794 = 1.88

For 60%,

qnorm(0.80)=0.8416212 = 0.84

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

Ans:- n= 25, n-1 =25-1 =24

For 95%,

qt (0.975,24)= 2.063899

For 96%,

qt (0.98,24)=2.171545

For 99%,

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:- Considering it as a normal distribution,

Given,

n = 18,

µ = 270 days,

= 260 days,

S = 90 days

t-score = -µ/(s/n^(1/2) ),

Putting all the values in above formula,

t-score= -0.471

df =n-1 = 18-1 =17

In R,

pt(tscore,df) = pt(-0.471, 17) = 0.321

Probability is 32%