|  |  |
| --- | --- |
| 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 | 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 | Interval |
| Religious Preference | Ordinal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Ordinal |

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

ANS:

HHT:

Probability = (1/2) \* (1/2) \* (1/2) = 1/8

HTH:

Probability = (1/2) \* (1/2) \* (1/2) = 1/8

THH:

Probability = (1/2) \* (1/2) \* (1/2) = 1/8

Now, add up the probabilities for all three cases because they are mutually exclusive events

Total Probability = (1/8) + (1/8) + (1/8) = 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:

1. There is no way that the probability that sum is equal to 1, because the minimum sum is atleast 2.
2. Sum = 2: (1, 1)

Sum = 3: (1, 2) (2, 1)

Sum = 4 (1, 3) (2, 2) (3, 1)

Total = 6

the probability that sum is Less than or equal to 4 is 6/36

1. Sum = 6: (1, 5) (2, 4) (3, 3) (4, 2) (5, 1) (6,6)

Total = 6

the probability that sum is Sum is divisible by 2and 3 is 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 (red) + 3 (green) + 2 (blue) = 7 balls

Sample space = 21

n(a) = 10 ( by using combination 5c2)

The probability that none of the balls drawn is blue = 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:

the Expected number of candies for a randomly selected child =(1 \* 0.015) + (4 \* 0.20) + (3 \* 0.65) + (5 \* 0.005) + (6 \* 0.01) + (2 \* 0.120)

The Expected number of candies = 3.08

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.

**Use Q7.csv file**

**ANS:**

Points: Mean = 3.596563, Median = 3.695000, Mode = 3.07,3.92 Variance = 0.285881633041, Standard Deviation = 0.534679, Range=2.17

Score­: Mean = 3.217250, Median = 3.325000, Mode = 3.44 Variance = 0.957378100849, Standard Deviation = 0.978457, Range = 3.9110

Weigh: Mean = 17.848750, Median = 17.710000, Mode = 17.02,18.90 Variance = 3.193165285249, Standard Deviation = 1.786943, Range = 8.3999

On Points The mean is less than median ,On Score The mean is less than median, On Points The mean is more than median.

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:

108+110+123+134+135+145+167+187+199=1308 pounds

Also there is 9 data points, then

1308/9 = 145.9999

So, the expected value of the weight of a randomly chosen patient from the clinic is approximately 145 pounds.

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Ans:**

**(Refer IPYNB)**

Skewness- Speed = -0.117510

Distance = 0.806895

Kurtosis -Speed = -0.508994

Distance = 0.405053

Inferences:

In Data,Speed has negative value which means Left- skewed , so the mean is less than median(The data is Nearly Symmetrical).

In Data, Distance has positive value which means Right-skewed, so the mean is more than median(The data is Slightly Skewed).

In Data, Speed has negative value which means Negative Kurtosis, it suggests lighter tails and a flatter distribution.

In Data, Distance has positive valuewhich means Positive Kurtosis, it indicates heavier tails and a more peaked distribution.

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Ans:**

**(Refer IPYNB)**

Skewness-Speed= 1.611450

Weight= -0.614753

Kurtosis- Speed= 2.977329

Weight= 0.950291

**Inferences:**

In Data, Speed has Positive value which means Right- skewed , so the mean is more than median(The data is Exetremely Skewed).

In Data, Weight has negative value which means Left-skewed, so the mean is less than median(The data is Slightly Skewed).

In Data, Speed has positive value which means Positive kurtosis, it indicates heavier tails and a more peaked distribution(Platykurtic).

In Data, Weight has positive value value which meansPositive kurtosis, it indicates heavier tails and a more peaked distribution(Mesokurtic).

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



**Inferences:**

In the histogram, there is less frequencies compare to other range, so it is left skewed.

In box plot, as we can see there is too many outliers in 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 Interval: (197.35, 202.65) pounds

98% Confidence Interval: (196.73, 203.27) pounds

96% Confidence Interval: (197.13, 202.87) pounds

**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, Varience=25.5294,

Standard Deviation=5.0526

1. What can we say about the student marks?

We can say that more of the students have similar marks

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

Ans: When the mean, median of data are equal then there is a minute skewness or no skewness in the data, Skewness is Asymmentric data which the mean and median are not same so skewness does not exist.

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

Ans: When the Mean is Greater than Median, Then it is Right skewed, which is positively distributed.

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

Ans: When the Mean is Less then Median(ie.median > mean), Then it is Left skewed, Which is negatively distributed.

Q16) What does positive kurtosis value indicates for adata ?

Ans: Positive kurtosis value indicates that distribution is peaked and possesses thick tails

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

Ans: Negative kurtosis value indicates that distribution is stretched and possesses thin tails

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



What can we say about the distribution of the data?

Ans: More then 50% of data lies on 10-18 range

What is nature of skewness of the data?

Ans: The Nature of the skewness is “Left skewed”

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

Ans: To find the IQR,

IQR= Upper Quartile – Lower Quartile

= 18 – 10

= 8(approx.)

Q19) Comment on the below Boxplot visualizations?



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

Inference:

When Compared to box plot 1 to box plot 2, we can say that box plot 1 has small data compared to box plot

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

* 1. P(MPG>38)
  2. P(MPG<40)
  3. P (20<MPG<50)

**ANS:**

To calculate the probability of MPG of cars we need to find the Mean&Standarad deviation

Mean=34.422076 Standard Deviation=9.131445

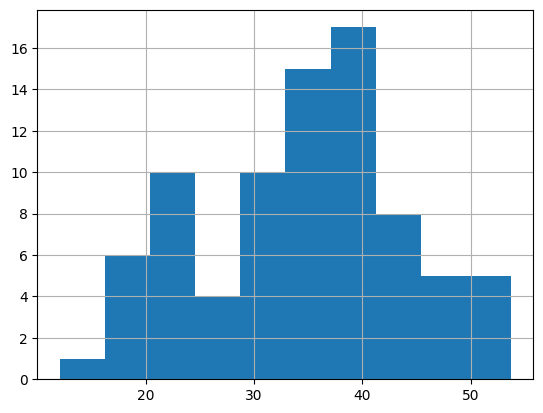
1. P(MPG>38)=0.34759392515827137
2. P(MPG<40)=0.7293498604157946
3. P (20<MPG<50)=0.8988689076273199

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

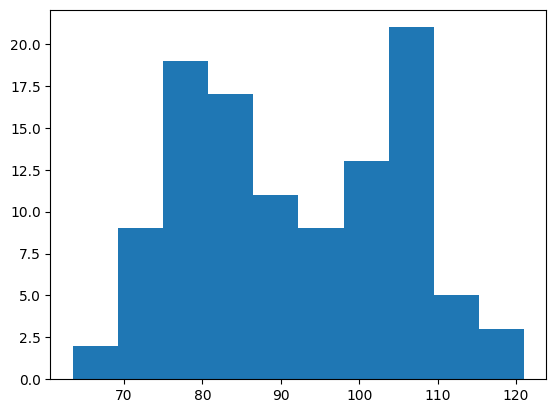
ANS:



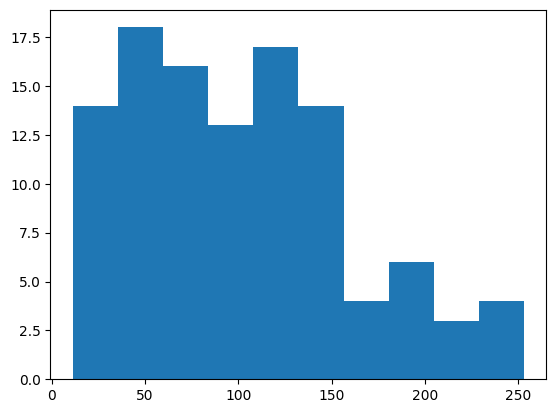
* By seeing the above histogram we can say that it isn’t bell shaped and also Mean(Mean=34.422076) is not equal to Median(Median=35.152727)
* Also by the empirical rule of Normal distribution we can say that the 68% of data is not within One Standard Deviation of the Mean

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 

* By seeing the above histogram we can say that it isn’t bell shaped and also Mean(Mean=91.901835) is not equal to Median(Median=90.800000)
* Also by the empirical rule of Normal distribution we can say that the 68% of data is not within One Standard Deviation of the Mean
* Hence we can say that the data does not follow normal distribution

Adipose tissue

* By seeing the above histogram we can say that it isn’t bell shaped and also Mean(Mean=101.894037) is not equal to Median(Median=96.540000)
* Also by the empirical rule of Normal distribution we can say that the 68% of data is not within One Standard Deviation of the Mean
* Hence we can say that the data does not follow normal distribution

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

ANS:

Z-score for a 90% confidence interval = 1.645 (approximately)

Z-score for a 94% confidence interval = 1.880 (approximately)

Z-score for a 60% confidence interval = 0.842 (approximately)

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

ANS:

Since n=25, and also we know that degrees of freedom(df) is (n-1)

df=24

|  |  |  |  |
| --- | --- | --- | --- |
| T scores of confidence interval(df=24) | t value for Right Tailed Probability | t value for Left Tailed Probability | t value for Two Tailed Probability |
| 95% | 0 | - 0 | ± 0.6849 |
| 96% | 0.2562 | - 0.2562 | ± 0.8569 |
| 99% | 1.3178 | -1.3178 | ± 1.7109 |

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:

n=18, population mean=270, sample mean=260

Standard deviation(sample)=90.

(population standard deviation is unknown)

Let us take,

ho>=260 & h1 < 260

we know that,

degrees of freedom = n-1 = 17

Level of significance(alpha)= 0.05%

Since we don’t know the population standard deviation, then we have to use t-statistics

By using the known codes for finding p value

P value= 0.32167253567098364

Level of significance(alpha)=0.05 < P value= 0.32167253567098364

So we accept null hypothesis(h0), hence we can say that bulb lasts more than 260 days