|  |  |
| --- | --- |
| 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 | 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 | Nominal |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | interval |
| SAT Scores | Ratio |
| Years of Education | Ratio |

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

Ans –

Sample space = {HHH, HHT, HTH, HTT, TTT, TTH, THT, THH}

Count of sample space = 8

Probability = No of favorable outcome / total no of outcome

Favorable outcome = {HHT, HTH, THH}, count = 3

Probability = 3/8 = 0.375 or 37.5%

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

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

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

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

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

Count = 36

Probability = No of favorable outcome / total no of outcome

1. Equal to 1

Ans – as two dices are being rolled the minimum sum would be 2

Favorable outcome = { }, count = 0

Probability = 0/36 = 0

1. Less than or equal to 4

Ans –

Favorable outcome = {(1,1), (1,2), (1,3), (2,1), (2,2), (3,1)}, count = 6

Probability = 6/36 = 1/6 = 0.166 = 16.6%

1. Sum is divisible by 2 and 3

Ans –

Favorable outcome = {(1,5), (2,4), (3,3), (4,2), (5,1), (6,6)}, count = 6

Probability = 6/36 = 1/6 = 0.166 = 16.6%

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 –

The total number of balls in the bag is 2 (red) + 3 (green) + 2 (blue) = 7.

Drawing one non-blue ball and one blue ball

The number of favorable outcomes for this scenario is the number of ways to choose 1 non-blue ball from the 2 red and 3 green balls, multiplied by the number of ways to choose 1 blue ball from the 2 blue balls.

Number of favorable outcomes = (C(2, 1) + C(3, 1)) \* C(2, 1) = (2 + 3) \* 2 = 1

Total number of possible outcomes = C(7, 2) = 21

Therefore, the total number of favorable outcomes is 4 + 10 = 14.

The probability of not drawing any blue balls is given by:

Probability = Number of Favorable Outcomes / Total Number of Possible Outcomes

Probability = 14 / 2

Simplifying the fraction, we get:

Probability = 2 / 3 = 0.66 = 66%

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 values when the probability is provided for a data point is given as

E(X) = ∑ X \* P(X)

|  |  |  |  |
| --- | --- | --- | --- |
| CHILD | Candies count | Probability | Expected Value |
| A | 1 | 0.015 | 1 \* 0.015 = 0.015 |
| B | 4 | 0.20 | 4 \* 0.20 = 0.80 |
| C | 3 | 0.65 | 3 \* 0.65 = 1.95 |
| D | 5 | 0.005 | 5 \* 0.005 = 0.025 |
| E | 6 | 0.01 | 6 \* 0.01 = 0.06 |
| F | 2 | 0.120 | 2 \* 0.120 = 0.240 |

To get the expected number of candies for a randomly selected child we sum up all the calculated expected values

0.015 + 0.80 + 1.95 + 0.025 + 0.06 + 0.240 = 3.09

the Expected number of candies for a randomly selected child is 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.

**Use Q7.csv file**

|  |  |  |  |
| --- | --- | --- | --- |
| Metrics | Points | Score | Weigh |
| Mean | 3.596 | 3.217 | 17.848 |
| Median | 3.695 | 3.325 | 17.71 |
| Mode | 0 3.07,  1 3.92 | 0 3.44 | 0 17.02  1 18.90 |
| Standard Deviation | 0.534 | 0.978 | 1.786 |
| Variance | 0.285 | 0.957 | 3.193 |
| range | 2.17 | 3.911 | 8.399 |

Ans –we get the idea about skewness when compare mean and median of columns of data, here we can see that mean and media are equivalent to each other with verry slight difference and mean, median, & mode are also equivalent to each other so we conclude that data is normally distributed

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 values when the probability is provided for a data point is given as

E(X) =∑ (X \* P(X))

Count of weight of patient at a clinic = 9

Probability of patient selected randomly = 1/9

E(X) = (108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199) \* 1/9

= 1308 \* 1/9

= 1308/9

= 145.33

the Expected Value of the Weight of that patient chosen randomly = 145.33

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

Cars speed and distance

Use Q9\_a.csv

Df[‘sp’].skew()

SP and Weight(WT)

Use Q9\_b.csv

Ans –

Skewness value of Speed column: -0.117

Kurtosis value of Speed column: -0.508

Speed column contains a very low negative skewness

Skewness value of Distance column: 0.806

Kurtosis value of Distance column: 0.405

Distance column contains very high positive skewness

Skewness value of SP column: 1.611

Kurtosis value of SP column: 2.977

SP column contains very high positive skewness

Skewness value of WT column: -0.614

Kurtosis value of WT column: 0.950

WT column contains very high positive skewness

Q10) Draw inferences about the following boxplot & histogram



Ans – From the given boxplot the we can observe that the bin from 50-100 has the the heighest frequeny (approx 200) that is Chicken weight ranging from 50-100 has the most number of occurance.

The data is heavily right or positve skewed as there is presence of outlier or extreme values on the upper extreme or right ends of the given Histogram after 300(approx)

Most of the data lies between 0 to 150.



Ans – There is presence of outlier on the upper extreme of the box plot, and due presence of outliers on upper extreme we can say that data is right or 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 –

94% confidence interval = 198.7376 - 201.2623

96% confidence interval = 198.6214, 201.378

98% confidence interval = 198.4381 - 201.5618

**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 = 41.0

Median = 40.5

Variance = 24.11

Standard Deviation = 4.910

1. What can we say about the student marks?

Ans – As mean is greater than median, we can say that data is slightly Right skewed or positive skewed and outlier lies towards right end or upper extreme end of the given data. 41 is the most occurring marks.

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

Ans – When mean = median the data is symmetric in Nature, there will be zero skewness i.e. skewness is not present in data.

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

Ans – When mean > median there will be right skewness, there is presence of positive skewness

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

Ans – When mean < median there will be Left skewness, there is presence of negative skewness

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

Ans – positive skewness indicates the data in Right skewed i.e. the peak of skewness is towards right side, the mean > median, outliers values are more towards upper extreme or right tail.

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

Ans –negative skewness indicates the data in Right skewed i.e. the peak of skewness is towards left side, the mean < median, outliers values are more towards lower extreme or left tail

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



What can we say about the distribution of the data?

Ans – The distribution of data is Asymmetric in nature as the whiskers on both end are not of equal lengths.

What is nature of skewness of the data?

Ans – The data heavily left or negative Skewed

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

Ans –

Q1 – 10 (Approx)

Q3 – 18 (approx)

IQR = Q3 – Q1

IQR = 18 – 10 (Substituting the values of Q1 & Q3)

IQR = 8 (Approximately)

Q19) Comment on the below Boxplot visualizations?



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

Ans – There is a normal distribution with median of both the data is same, there is difference in upper , lower extreme & IQR values both the 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

Ans –

P(MPG>38) = 0.34759392515827137

P(MPG<40) = 0.7293498762151609

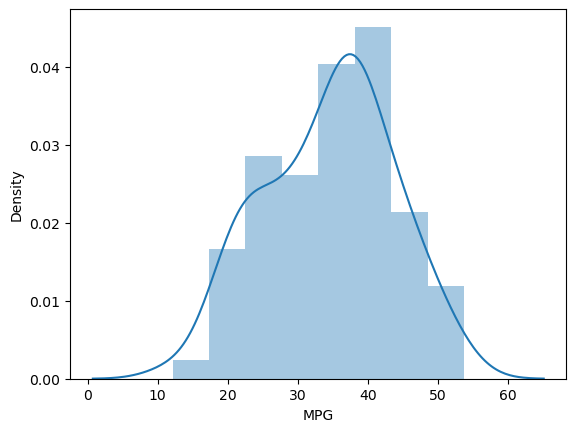
P (20<MPG<50) = 0.8988689169682047

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans –

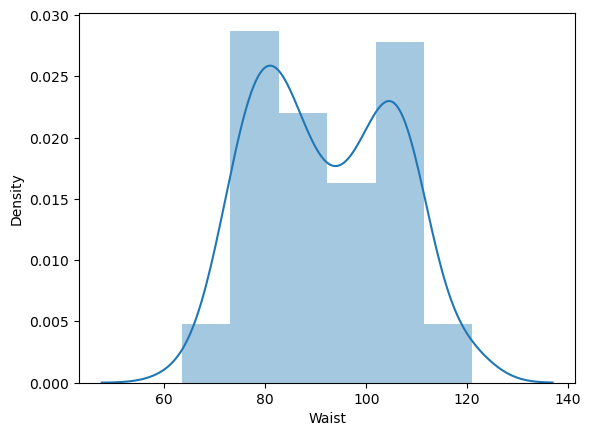


Skewness of MPG column = -0.1779

From the above visualization we can see that the curve of density graph is nearly bell shaped and as skewness is near to 0 we can say that MPG follows normal distribution.

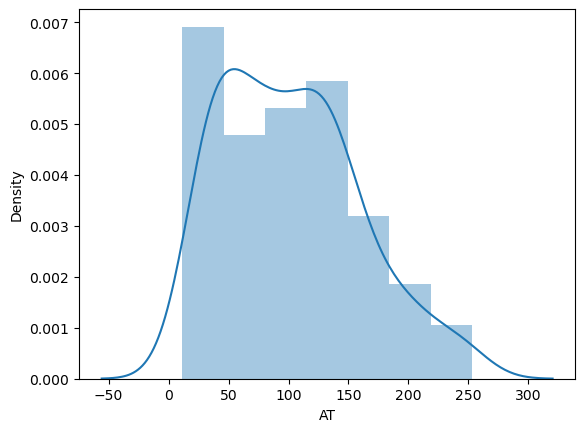
1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution

Dataset: wc-at.csv



Skewness of Waist 0.134056

For waist from the above visualization, we can see that the curve of density graph is nearly bell shaped and as skewness is near to 0 we can say that waist follows normal distribution



Skewness AT 0.584869

For Adipose Tissue AT from the above visualization, we can see that the curve of density graph is not a bell shaped and as skewness ranges between -0.5 to +0.5 we can say that Adipose Tissue AT is moderately skewed and does not follow normal distribution

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

Ans –

Z - score at 60 % confidence interval - 0.2533

Z - score at 90 % confidence interval - 1.2815

Z - score at 94 % confidence interval - 1.5547

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

Ans –

T score at 95% of confidence interval with degree of freedom 24 = 1.71088

T score at 96% of confidence interval with degree of freedom 24 = 1.82805

T score at 99% of confidence interval with degree of freedom 24 = 2.49215

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 –

0.3216 is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days