|  |  |
| --- | --- |
| 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 | Ratio |
| Time on a Clock with Hands | Nominal |
| 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.** The sample space for tossing 3 coins will be ( TTT, TTH, THT, THH, HTT, HTH, HHT, HHH )

Hence the probability of the Event E (Two heads and one tail) will be :

P(E) = Number of favourable outcomes of E/Total number of possible outcomes in sample space

= 3/8 (THH, HTH, HHT)

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 2 and 3

**Ans.** The sample space for rolling two dice will be

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

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

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

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

1. Let A be the event of getting sum equal to 1.

P(A) = Number of favourable outcomes of A/Total number of outcomes in Sample space

P(A)= 0/36

P(A)= 0

1. Let B be the event of getting sum less than or equal to 4

P(B) = Number of favourable outcomes of B/Total number of outcomes in Sample space

B = { (1,1) , (1,2) , (1,3) , (2,1) , (2,2) , (3,1) }

P(B) = 6/36

P(B) = 1/6

1. Let C be the event that sum is divisible by 2 and 3.

P(C) = Number of favourable outcomes of C/Total number of outcomes in sample space

C = { (1,5) , (2,4) , (3,3) , (4,2) , (5,1) , (6,6) }

P(C) = 6/36

P(C) = 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 no of balls = 2 Red + 3 Green + 2 Blue

= 7 Balls

Let S be an event that two balls are drawn at random

N(S) = 7C2

N(S) = 7X6/2X1

N(S) = 21

Let A be an event that none of the ball drawn is blue

N(A) = 5C2

= 5X4/2X1

= 10

P(A)(Probability of A) = N(A)/N(S) = 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.** To calculate the expected no of candies for a randomly selected child, we need to multiply each candy by count by its corresponding probability.

Expected Value = ∑ ( probability \* Value)

∑P(x).E(x)

The expected no. of candies(E) can be calculated as:

E = (1\*0.015) + (4\*0.20) + (3\*0.65) + (5\*0.005) + (6\*0.01) + (2\*0.120)

E = 3.09

The expected number of candies for a randomly selected child is approximately 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**

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 = ∑( probability \* Value)

∑P(x).E(x)

There are 0 patients, so the probability of selecting each patient randomly = 1/9

E(x) 108, 110, 123, 134, 135, 145, 167, 187, 199

P(x) 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9, 1/9

Expected Value = (1/9)(108) + (1/9)(110) + (1/9)(123) + (1/9)(134) + (1/9)(135) + (1/9)(145) + (1/9)(167) + (1/9)(187) + (1/9)(199)

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

=(1/9)(1308)

=145.33

OUTPUT: The expected value of the weight for a randomly chosen patient is approximately 145.33 pounds.

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

**Cars speed and distance**

**Use Q9\_a.csv**

**SP and Weight(WT)**

**Use Q9\_b.csv**

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



Ans. Inference: 50-100 weight having more frequency i.e. 180

550-400 weight have very little frequency 5

Data is right skewed or Positively Skewed.

Data is not normally distributed.

In right-skewed data Mode<Mean<Median



Ans. 7 Outliers are present in the above box plot

Positive Skewness .i.e data is right-skewed

Data is not normally distributed.

Most of the data values are clustered near the median, i.e. they are scattered.

**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?

**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?

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

Ans. The Skewness value generally ranges between -1 to 1, but in case when mean and median are equal the skewness value is 0(zero) or almost equivalent to 0.The data is normally distributed

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

Ans. The Skewness value is more towards the left-hand side of 0 i.e. it is negative because the data is left Skewed or Negatively Skewed, which means the data is more on the right-hand side than on the left-hand side.

Skewness is a measure of the asymmetric of a distribution.

When the mean is greater than the median, it implies that the distribution has a longer tail on the left side.

A negative skewness value indicates that the tail of the distribution is skewed to the left, there are more extreme values on the right side distribution

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

Ans. The Skewness value is more towards the right-hand side of 0 i.e. it is positive because the data is right Skewed or Positively Skewed, which means the data is more on the left-hand side than on the right-hand side.

Skewness is a measure of the asymmetric of a distribution.

When the median is greater than the mean, it implies that the distribution has a longer tail on the right side.

A Positive skewness value indicates that the tail of the distribution is skewed to the right, there are more extreme values on the left side distribution

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

Ans. Positive kurtosis is commonly known as ‘’Lepokurtic”

Positive values of kurtosis indicate that distribution is peaked and possesses heavier tails and is concentrated more on the central peak compared to a normal distribution,it suggests that data has more outliers (or)extreme values

Kurtosis is the measure of the tailedness of distribution and peak relatively to the normal distribution

A positive kurtosis value indicates the dataset has heavier tails and more peaked central peak, suggesting a higher likelihood of extreme values (or)outliers compared to a normal distribution

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

Ans. Negative kurtosis is commonly known as “platykurtic”

Negative values of kurtosis indicate that distribution is peaked and possesses light tails and is concentrated less on the central peak compared to a normal distribution,it suggests that data has fewer outliers (or)extreme values

Kurtosis is the measure of the tailedness of distribution and peak relative to the normal distribution

A negative kurtosis value indicates dataset has lighter tails and less peaked central peak, suggesting a lower likelihood of extreme values (or)outliers compared to a normal distribution

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



What can we say about the distribution of the data?

Ans. In the boxplot the distribution of the data is “Left side (or) Negatively skewed data”, we can see so many outliers (or) extreme values

What is nature of skewness of the data?

Ans. In the boxplot visualization move towards the left skewed so, it is a “Negatively skewed data”

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

Ans. IQR stands for – “Inter Quartile Range”

Inter Quartile Range= Q3-Q1  
Q3 is the last value in the boxplot

Q1 is the first value in the boxplot

OUTPUT= 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 Boxplot 1 shows that most of the values are clustered near the median, i.e. they are not scattered.

The Boxplot 2 indicates that the values of the data are distributed and scattered i.e. they are not clustered around the median.

Q1 is 25%, Q3=75%.IQR is 50% for both the boxplots. So we can say both the distributions follow normal distribution i.e. Mean=Median=Mode.

There are no outliers in both the Boxplots.

The Variance of the Boxplot 1 is very less as compared to the Boxplot 2.

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)

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

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

Dataset: wc-at.csv

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

Ans. We need to calculate the Z-score for 90% confidence interval from the Z-score table

90% confidence interval =1.645 (from Z-score table we will get these values)

60% confidence interval =0.842 (from Z-score table we will get these values)

94% confidence interval for that we need to find Area

Area= 1+ confidence interval/2 =(1 + 94 /100) /2

= 1 + 0.94 / 2

=0.9700

We need to find the 0.9700 in the Z- score table

=1.8+0.08

OUTPUT = 1.88 (approximately)

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

Ans. Sample size(n)= 25

Degree of freedom =sample size -1

=24

From the t- score table ,we need to find the 95% confidence interval

95 % confidence interval =2.064

99 % confidence interval=2.797

96 % confidence interval=2.171

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