|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete data |
| Results of rolling a dice | Discrete data |
| Weight of a person | Continuous data |
| Weight of Gold | Continuous data |
| Distance between two places | Continuous data |
| Length of a leaf | Continuous data |
| Dog's weight | Continuous data |
| Blue Color | Discrete data |
| Number of kids | Discrete data |
| Number of tickets in Indian railways | Discrete data |
| Number of times married | Discrete data |
| Gender (Male or Female) | Discrete data |

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 data |
| High School Class Ranking | Ordinal data |
| Celsius Temperature | Interval data |
| Weight | Ratio data |
| Hair Color | Nominal data |
| Socioeconomic Status | Ordinal data |
| Fahrenheit Temperature | Interval data |
| Height | Ratio data |
| Type of living accommodation | Nominal data |
| Level of Agreement | Ordinal data |
| IQ(Intelligence Scale) | Interval data |
| Sales Figures | Ratio data |
| Blood Group | Nominal data |
| Time Of Day | Intervel data |
| Time on a Clock with Hands | Interval data |
| Number of Children | Nominal data |
| Religious Preference | Nominal data |
| Barometer Pressure | Interval data |
| SAT Scores | Interval data |
| Years of Education | Ratio data |

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

**SOL :**

when three coins tossed possible outcomes :

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

P(2 heads and 1 tail) = {THH},{HTH},{HHT}

=1/8+1/8+1/8

ANS: P(HH&T) =3/8 or 0.375

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

a)Equal to 1

**SOL:**

If two dices were rolled, then total possible cases = 36

Total Favourable cases (Having sum =1) = 0

As minimum sum is 2 for outcome (1,1)

**P(sum =1) = 0.**

Hence, probability is

1. Less than or equal to 4

**SOL:**

If two dices were rolled, then total possible cases = 36

Total Favourable cases = (1,1) (1,2) (1,3) (2,1) (2,2) (3,1)

**Probability = Number Of Favorable outcomes**

**Number Of Possible Outcomes**

= 6/36

ANS: P(≤4) =1/6

c) Sum is divisible by 2 and 3

**SOL:**

**When two dice are rolled, sample space is given as:**

(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)

Therefore,

**Total number of possible outcomes = 36**

Favorable outcomes = sum is divisible by 2 and 3

Sum should be divisible by both 2 and 3

Favorable outcomes = (1 , 5) , (3 , 3) , (4 , 2) , (5 , 1) , (6 , 6) ,(2 , 4)

Therefore,

Number of favorable outcomes = 6

**The probability of an event is given as:**

Probability = Number Of Favorable Outcomes

Number Of Possible Outcomes

= 6/36

=1/6

Thus the probability that sum is divisible by 2 and 3  is

ANS = 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?

**SOL:**

Total number of balls = 2R + 3G + 2B

= 7 balls  
Let S be the sample space.  
 n(S) = Number of ways of drawing 2 balls out of 7  
 =7C2​ =(7x6) / (2x1)  
 n(S)=21  
 E = Event of drawing 2 balls, none of which is blue.

n(E) = Number of ways of drawing 2 balls out of (2 + 3) balls.  
 =5C2​=(5x4) / (2x1)  
 n(E)=10

ANS : P(E)=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

**SOL:**

Expected number of candies for a randomly selected child

= 1 \* 0.015 + 4\*0.20 + 3 0.65 + 50.005 + 6 \*0.01 + 2 \* 0.12

= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

= 3.090

The expected number of candies for 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**

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?

Expected Value = ∑ (probability \* Value)

= ∑ P(x). E(x) There are 9 patients,

Probability of selecting each patient = 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)x(1308)

= 145.33

Expected Value of the Weight of that patient = 145.33

**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**



**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?

**Solution:**

The information given is: • Sample mean of . • Sample standard deviation of . • Sample size of . The interval is:

• In which t is the critical value for the two-tailed confidence interval. Considering a 94% confidence level, using a calculator, with 200 - 1 = 199 df, the critical value is t = 1.8916, hence:

The 94% confidence interval is (198.73, 201.27). Considering a 96% confidence level, using a calculator, with 200 - 1 = 199 df, the critical value is t = 2.0673, hence:

The 96% confidence interval is (198.61, 201.39). Considering a 98% confidence level, using a calculator, with 200 - 1 = 199 df, the critical value is t = 2.3452, hence:

The 98% confidence interval is (198.43, 201.57).

The 94% confidence interval is (198.73, 201.27). Considering a 96% confidence level, using a calculator, with 200 - 1 = 199 df, the critical value is t = 2.0673, hence:

The 96% confidence interval is (198.61, 201.39). Considering a 98% confidence level, using a calculator, with 200 - 1 = 199 df, the critical value is t = 2.3452, hence:

The 98% confidence interval is (198.43, 201.57).

**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.

Mean: As we know,

Mean= Sum of the terms/ Number of the terms.

**Mean**=(34+36+36+38+38+39+39+40+40+41+41+41+41+42+45+49+56)/18=738/18=41

**Median:** For median first I have to arrange in ascending order but here scores are already in ascending order so I directly calculated the median. Median = (9th+ 10th Term)/2

= (40+41)/2

= 40.5

**Variance:** Mean(m)=41 Scores(s) s-m (s-m)^2 34 -7 49 36 -5 25 36 -5 25 38 -3 9 38 -3 9 39 -2 4 39 -2 4 40 -1 1 40 -1 1 41 0 0 41 0 0 41 0 0 41 0 0 42 1 1 45 4 16 49 8 64 56 15 225 Sum 0 433

Variance = 433/17= 25.47

**standard deviation:**

As We know standard deviation= (Variance)^(1/2) So standard deviation= (24.05) ^ (1/2) = 5.05

1. What can we say about the student marks?

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

**Answer:**

Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed. Skewness can be quantified as a representation of the extent to which a given distribution varies from a normal distribution. A normal distribution has a skew of zero. When mean, median and mode is equal then it is a normal distribution. So, the nature of skewness is zero when mean and median of data is equal.

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

**Answer:**

Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed. A positively skewed distribution is the distribution with the tail on its right side. The value of skewness for a positively skewed distribution is greater than zero. When the value of mean is greater than median and mode then it is called positive skewed. So the nature of skewness is positive when mean > median.

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

**Answer:**

Skewness refers to a distortion or asymmetry that deviates from the symmetrical bell curve, or normal distribution, in a set of data. If the curve is shifted to the left or to the right, it is said to be skewed. A negatively skewed distribution is the distribution with the tail on its left side. The value of skewness for a negatively skewed distribution is less than zero. When the value of mean is less than median and mode then it is called negative skewed. So the nature of skewness is negative when median > mean.

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

**Answer:**

Kurtosis refers to the degree of presence of outliers in the distribution. It is a statistical measure, whether the data is heavy-tailed or light-tailed in a normal distribution. The excess kurtosis is used in statistics and probability theory to compare the kurtosis coefficient with that normal distribution. It can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near to zero (Mesokurtic distribution). Since normal distributions have a kurtosis of 3, excess kurtosis is calculating by subtracting kurtosis by 3. Excess kurtosis = Kurt – 3 Leptokurtic is having very long and skinny tails, which means there are more chances of outliers. Positive values of kurtosis indicate that distribution is peaked and possesses thick tails. An extreme positive kurtosis indicates a distribution where more of the numbers are located in the tails of the distribution instead of around the mean.

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

**Answer:**

Kurtosis refers to the degree of presence of outliers in the distribution. It is a statistical measure, whether the data is heavy-tailed or light-tailed in a normal distribution. The excess kurtosis is used in statistics and probability theory to compare the kurtosis coefficient with that normal distribution. It can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near to zero (Mesokurtic distribution). Since normal distributions have a kurtosis of 3, excess kurtosis is calculating by subtracting kurtosis by 3. Excess kurtosis = Kurt – 3 Platykurtic having a lower tail and stretched around center tails means most of the data points are present in high proximity with mean. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution. A distribution with a negative kurtosis value indicates that the distribution has lighter tails than the normal distribution

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



What can we say about the distribution of the data?

What is nature of skewness of the data?

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

Q19) Comment on the below Boxplot visualizations?



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

**solution:**

For 90% confidence interval: We have the significance level at 5 % (as it is a two tailed test) that is: α = 5 % = 0.05 z at α = 0.05 from the z table will be: z = 1.645. For 94 % confidence interval, we get: We have the significance level at 3 % (as it is a two tailed test) that is: α = 3 % = 0.03 z at α = 0.03 from the z table will be: z = 1.555. For 60 % confidence interval, we get: We have the significance level at 20 % (as it is a two tailed test) that is: α =20 % = 0.2 z at α = 0.2 from the z table will be: z = 0.253 Therefore, we get that the z score at 90% confidence interval is 1.645, at 94% confidence interval is 1.555 and at 60% confidence interval is 0.253

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

: a)The sample size is n=25, So the degrees of freedom is n−1=25−1=24 Thus, we are interested in the quantity t(α/2)=t(0.05/2)=t(0.025) for a t-distribution with 24 degrees of freedom. Upon using a t-table, we see that the critical t-value for this 95% confidence interval is t(α/2)=2.064.

b)Upon using a t-table, we see that the critical t-value for this 96% confidence interval is t(α/2)=2.164. c)Upon using a t-table, we see that the critical t-value for this 99% confidence interval is t(α/2)=2.797.

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