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

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

Solution: -

Binomial distribution

3 – tails 8 – Outcomes

T T H What is the probability getting 0 tails

T T H P (0) =1/8

T H T P (1) =3/8

T H H P (2) =3/8

H T T P (3) =1/8

H T H

H H T

H H H

Total number of ways = 2\*2\*2=8

Favorable cases=7

P (A) =7/8

OR

Probability of getting at least one head =1-P (number of head)

= 1-(1/8)

= 7/8

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 2 and 3

Solution: -

1. The sample space ‘S’ of two dice is shown below.

S = {(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)}

P(X) = Favorable events

Total number of possible outcomes

Let ‘E’ be the event “sum equal to 1”, there are no outcomes which correspond to a sum equal to 1. Hence,

P(E) = n(E)/n(S) = 0/36 =0

1. Three possible outcomes give a sum equal to 4.

E = (1,3), (2,2), (3,1), hence,

P(E) = n(E)/n(S) = 3/36 = 1/12

n(B) = 18/36

1. Sum is divisible 2 and 3

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

n(A) =12; n(A)

n(S)

= 12/36

B = {(1,1), (1,3), (2,1), (2,2), (2,4), (2,6), (3,1), (3,3), (3,5), (4,2), (4,4), (4,6)(5,1), (5,3), (5,5), (6,2), (6,4), (6,6)

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?

Solution: -

Total number of balls = (2+3+2) = 1

Let ‘S’ be the sample space.

Then, n(S) =Number of ways of drawing 2 balls out of 7 = 7C2 =21

Let, 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 = 10

Therefore, P(E) =n(E)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

Solution: -

Expected number of candies for a randomly selected child = 3.09

Expected number of candies for a randomly selected child.

= 1\*0.015+4\*0.20+3\*0.65+5\*0.005+6\*0.001+2\*0.12

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

= 3.090

Expected number of candies for a randomly selected child = 3.09

Q7) Calculate mean, median, mode, variance, standard deviation, range & comment about the values /draw inferences, for the given data set.

* 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**

Solution: -

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| Mean | 3.596 | 3.22 | 17.85 |
| Median | 3.695 | 3.33 | 17.71 |
| Variance | 0.285 | 0.96 | 3.19 |
| Standard deviation | 0.534 | 0.98 | 1.79 |
| Range | 2.76-4.93 | 1.513-5.424 | 14.5-22.9 |

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?

Solution: -

Expected value = (Probability \*value)

∑ P(X)\*P(X)

There are 9 participants.

Probability of selecting each patient = 9

EX: - 108, 110, 123, 134, 145, 167, 187, 199

P(X) 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

|  |
| --- |
| 145.33 |

Expected value of the weight of the patient =

**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**

Solution: -

Skewness of speed = -0.117510

Skewness of distance = 0.806895

Kurtosis of speed = -0.508994

Kurtosis of distance = 0.405053

Skewness of SP = 1.611450

Skewness of WT = -0.614753

Kurtosis of SP =2.977329

Kurtosis of WT = 0.950291

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



Solution: -

1. The distribution is right skewed. Mean>median.
2. The above box plot 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?

Solution: -

1. Confidence interval = 94 % ()
2. Confidence interval = 98 % (µ0=2.05< \* < µ0 +2.05)
3. Confidence interval =96% (µ0 =1.75 < \* < µ0c+1.75)

Step-by-step explanation:

Sample size n= 300000

Sample mean X =200

Standard deviation S = 30

From z-table values of z(c)

CI 94 % Confidence level α = 6%, α = 0.06, z(c) = 1.55

CI 98% Confidence level α = 2%, α = 0.02, z(c) = 2.05

CI 96% Confidence level α = 4%, α =0.04, z(c) = 1.75

MOE = Z(C) \*σ/

1. MOE = 1.55\*30

=1.04

1. MOE = 2.05\*30

= 1.38

1. MOE = 1.75\*30

= 1.17

Then confidence interval

1. CI = 94% (µ0-MOE < \* < µ0-MOE)

CI = (

1. CI = 98%

CI = (µ0-2.05 < \* < µ0 +2.05)

1. CI = 96%

CI = (µ0 – 1.75 < \* < µ0 +1.75)

**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?

Solution: -

1. 34, 36, 36, 38, 39, 39, 40, 40, 41, 41, 41, 41, 42, 42, 45, 49, 56

n = 17

mean = 34+36+36+38+39+39+40+40+41+41+41+41+42+42+45+49+56

17

|  |
| --- |
| Mean = 41.17 = µ |

Median = 34, 36, 36, 38, 39, 39, 40, 40, 41, 41, 41, 41, 42, 42, 45, 49, 56

|  |
| --- |
| Median = 41 |

Variance = ∑(X - X BAR)2

n

= 170.09

17

= 10.0057

Standard deviation = σ = 2

n

=

1. Mean = µ = 41.17

Median =41

Variance = σ = 10.0057

Standard deviation = σ2 =3.1631

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

Solution: -

1. If the distribution is symmetric, then the mean is equal to the median, and the distribution has zero skewness. If the distribution is both symmetric and unimodal, then the mean = median mode.
2. If the mean, median, and the mode of a set of numbers are equal, it means, the distribution is symmetric. The more skewed is the distribution, greater is the difference between the median and mean, and we should lay greater emphasis on using the median as opposed to the mean.

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

Solution: -

If the mean is greater than the mode, the distribution is positively skewed. If the mean is less than the mode, the distribution is negatively skewed. If the mean is greater than the median, the distribution is positively skewed. If the mean is less than the median, and the distribution is negatively skewed.

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

Solution: -

If the mean is greater than the median, the distribution is positively skewed. If the mean is less than the median, the distribution is negatively skewed. If the median is greater than the mean on a set of test scores, describe the situation. The official answer is that the data are “skewed to the left”, with a long tail of low scores pulling the mean down more than the median.

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

Solution: -

Positive values of kurtosis indicates that a distribution is peaked and possess thick tails. An extreme positive kurtosis indicates a distribution where more of the values are located in the tails of the distribution rather than around the mean.

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

Solution: -

Negative values of kurtosis indicates that a distribution is flat and has thin tails. Platykurtic distributions are have negative kurtosis values. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution, with fewer values in its shorter (i.e., lighter and thinner) tails.

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)?   
Solution: -

Minimum value = 2

Maximum value =18

Q1: First quartile = 10

Q2: Second quartile or median = 15

Q3: Third quartile = 18

1. Each quartile has approximately 25% of the data.
2. The spreads of the four quarters are 10-2 = 8 (first quarter)

15-10 = 5 (second quarter)

18-15 = 3 (third quarter)

So, the third quarter has the smallest spread and the first quarter has the largest spread.

1. Range = maximum value – minimum value

= 18 – 2

= 16

1. Interquartile range: IQR = 18 – 10 = 8
2. The middle 50% (middle half) of the data has a range of 8 inches.

I take that its negatively skewed because the right quartile is the farther from the median.

>> Left skewed.

>> Nature of skewness 🡪 negative.

>> The interval 2-10 has more than 50% of the data so it has more data in it than the interval 15 through 18 which has 25% of the data.

Q19) Comment on the below Boxplot visualizations?



Solution: -

|  |  |
| --- | --- |
|  |  |

(Q1 – 1.5\*IQR) 25TH Percentile 75th Percentile (Q3+1.5\*IQR)

1. Inter quartile range is the range between upper quartile Q3 and lower quartile.

IQR = 250-275 = 25

IQR = 225-300 = 75

* 25% of the data lies in fig1.
* 75% of the data lies in fig2.

1. The above boxplot with respect to boxplot 1 and 2 are symmetrical.
2. Both are the normally distributed 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

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

Solution: -

1. P(MPG>38)

Mean (MPG) = 34.422

Mean (MPG) = 9.131

PNorm = 0.6525

1-0.65 = 0.35

P(MPG) = 0.35

1. P(MPG<40)

Pnorm = 0.72945

P(MPG<40) = 0.73

1. P(20<MPG<50)

Pnorm = (50, 34.42, 9.13)-(20, 34.42, 9.13)

= 0.89891

P(20<MPG<50) =0.898

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Solution: -

By using cars.csv file MPG of cars, QQ plot has normally distributed.

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

Dataset: wc-at.csv

Solution: -

By using wc-at.csv file Adipose tissue (AT) and waist circumference (waist),

QQ plot has normally distributed.

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

Solution:-

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SI NO | Confidence level | 90% | 94% | 60% |
| 1. | Significance level  (α = 1-confidencelevel) | = 1-0.90  = 0.06 | = 1-0.94  = 0.06 | = 1-0.60  = 0.4 |
| 2. | Significance level(α/2) | = 0.1/2  = 0.05 | = 0.06/2  = 0.03 | = 0.4/2  = 0.2 |
| 3. | Value to look up  In z-table | = 1-0.05  = 0.95 | = 1-0.03  = 0.97 | = 1-0.2  = 0.8 |
| 4. | Z- score | =1.645  From z-table | = 1.89  From z-table | = 0.845  From z-table |

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

Solution:-

Degree of freedom = n-1

= 25-1

df = 24

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| SI NO | Confidence level | 95% | 96% | 99% |
| 1. | Significance level (α=1-confidence level) | = 1-0.95  = 0.05 | = 1-0.96  = 0.04 | = 1-0.99  = 0.01 |
| 2. | Significance level(α/2) | = 0.05/2  = 0.025 | = 0.04/2  = 0.02 | = 0.01/2  = 0.005 |
| 3. | Value to look up in t-table. | = 1-0.025  = 0.975 | = 1-0.02  = 0.98 | = 1-0.005  = 0.995 |
| 4. | t-score | = 2.064  From z-table. | Table-B  = 2.172  From z-table. | = 2.797  From z-table. |

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

Solution: -

t-statistics for the data is given as follow.

X = Mean of the sample bulb = 260.

µ = Population mean = 270.

S = Standard deviation of the sample =90.

n = Number of items in the sample = 18.

t = = = 0.471

t = 0.471

For probability calculations the number of degrees of freedom is n-1, so here you need the t-distribution with 17 degrees of freedom.

The probability that t < -0.471 with 17 degrees of freedom assuming the population mean is true. The t-value is less than the t-value obtained with 17 degrees of freedom and a t score of -0.471, the probability of the bulbs lasting less than 260 days on average of 0.3218 assuming the mean the life of the bulbs is 300 days.