|  |  |
| --- | --- |
| 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 | Nominal |
| Number of kids | Discrete |
| Number of tickets in Indian railways | Discrete |
| Number of times married | Discrete |
| Gender (Male or Female) | Nominal |

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

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

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

n(S) = 8

p(2heads and 1 tails) = 3/8 = 0.375

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

Sol: 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)}

n(S) = 36

a)🡪 p(sum = 1) = 0/36 = 0

b)🡪 p(sum <= 4) = 6/36 = 0.1666

c)🡪 p(sum is divisible by 2 and 3) =6/36 = 0.1666

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 in a bag = 7

n(S) = 7

Two balls are drawn at random from bag i.e 7C2 = 21

n(2 balls random from bag) = 21

Number of ways of drawing 2 balls such that none is blue = Number of ways of drawing 2 balls from 2 red and 3 green balls

ie 5C2 = 10

p(none of the ball drawn is blue) = 5C2/7C2 = 10/21 = 0.471

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: Fomula for calculating the Expected value is *EV*=∑*P*(*Xi*​)×*Xi*​​

Expected value = 1\*0.015 + 4\*0.20 + 3\*0.65 + 5\*0.005 + 6\*0.01 + 2\*0.120

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

= 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.

Sol:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Mean | Median | Mode | Variance | Std Dev | Range |
| Points | 3.596 | 3.695 | 3.07 and 3.92 | 0.285 | 0.534 | 2.17 |
| Score | 3.217 | 3.325 | 3.44 | 0.956 | 0.978 | 3.911 |
| Weigh | 17.848 | 17.71 | 17.02,18.90 | 3.189 | 1.786 | 8.4 |

🡪The mean and median of features Points ,score and weigh tells that there is no outliers present in the dataset

🡪The variance and standard deviation is low for Points and Score which means the predictions will be done accurately

**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?

Sol: There are 9 patients

Probability of selecting each patient = 1/9

Expected Value = 108\*1/9 + 110\*1/9 + 123\*1/9 + 134\*1/9 + 145\*1/9 +

167\*1/9 + 187\*1/9 + 199\*1/9

= 145.33

Expected value of one of the patient is 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**Sol:**

|  |  |  |
| --- | --- | --- |
|  | **Skewness** | **Kurtosis** |
| **Speed** | **-0.117** | **-0.508** |
| **Distance** | **0.806** | **0.405** |

**🡪The skewness of speed is negative and less than 0 that means it is negatively skewed data**

**🡪The skewness of distance is positive and greater than 0 that means it is positively skewed data**

**🡪The kurtosis of speed is less than 0 that means it is platykurtic distribution**

**🡪The kurtosis of distance is more than 0 that means it is leptokurtic distribution**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Sol:**

|  |  |  |
| --- | --- | --- |
|  | **Skewness** | **Kurtosis** |
| **SP** | 1.611 | 2.977 |
| **WT** | -0.614 | 0.950 |

**🡪 The skewness of SP is positive and greater than 0 that means it is positively skewed data**

**🡪 The skewness of MT is negative and less than 0 that means it is negatively skewed data**

**🡪 The Excess kurtosis of SP is greater than 0 that means it is leptokurtic distribution**

**🡪 The Excess kurtosis of WT is close to 0 that means it is moderately mesokurtic distribution**

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



🡪The histogram is showing that the distribution of data is right skewed or positively skewed

🡪 The mean value of data is higher than the median and mode



🡪 The Boxplot is showing that there are lots of outliers are present at the higher end of the data.

🡪 The Median line (Q2) is close to the Q1 that means the data right skewed distributed.

**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?

Sol: 🡪Confidence Interval for 94% = x̄ ± z(σ/√n)

= [200-1.96(30/2000\*\*0.5) , 200+1.96(30/2000\*\*0.5)]

= [198.738, 201.261]

🡪Confidence Interval for 98% = x̄ ± z(σ/√n)

=[200-2.326(30/2000\*\*0.5) , 200+2.326(30/2000\*\*0.5)]

**= [**198.43 , 201.56]

🡪Confidence Interval for 98% = x̄ ± z(σ/√n)

=[200-2.58(30/2000\*\*0.5) , 200+2.58(30/2000\*\*0.5)]

=[198.62, 201.37]

**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?

Sol: Mean = 41

Median = 40.5

Variance = 25.529

Standard Deviation = 5.052

🡪 The student marks are normally distributed

🡪 No outliers are present in the students marks

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

Sol: when mean=median=0 then it is zero skewed distribution or normal distribution

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

Sol: when mean>median then it is positive skewed or right skewed distribution

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

Sol: when median>mean then it is negative skewed or left skewed distribution

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

Sol: Positive kurtosis value indicates that the distribution of data is leptokurtic

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

Sol: Negative kurtosis value indicates that the distribution of data is platykurtic

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



What can we say about the distribution of the data?

🡪Arround 25% of the data are less than 10 and 75% of the data are less than 18

🡪 The distribution of data is slightly negatively skewed or left skewed

What is nature of skewness of the data?

🡪 The nature of skewness of the data is left skewed or negative skewed

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

🡪 According to box plot the value of Q1 = 10, Q3 = 18

So the IQR would be IQR = Q3-Q1 = 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.

Sol:

|  |  |  |
| --- | --- | --- |
|  | Boxplot 1 | Boxplot 2 |
| Minimum value | ~240 | ~190 |
| Q1 | ~238 | ~225 |
| Q2 | ~262 | ~262 |
| Q3 | ~281 | ~312 |
| Maximum value | ~287 | ~337 |
| Distribution | Normal Distribution | Normal Distribution |
| Outliers | No outliers present | No outliers present |

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)

Sol:

🡪 a) n(MPG) = 81

n(MPG>38) = 33

P(MPG>38) = 33/81 = 0.40

🡪 b) n(MPG) = 81

n(MPG<40) = 61

P(MPG<40) = 61/81 = 0.75

🡪 c) n(MPG) = 81

n(20<MPG<50) = 69

P(20<MPG<50) = 69/81 = 0.85

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

🡪 The MPG of Cars follows fairly 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

Sol:🡪 Waist Circumference is fairly normal distribution.

🡪 Adipose Tissue is slightly positively skewed.

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

Sol:

|  |  |  |  |
| --- | --- | --- | --- |
|  | 90% CI | 94% CI | 60% CI |
| z-scores | 1.2815515655446004 | 1.5547735945968535 | 0.2533471031357997 |

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

Sol:

|  |  |  |  |
| --- | --- | --- | --- |
|  | 95% CI | 96% CI | 99% CI |
| t-scores | 2.063898561628020 | 2.171544676008067 | 2.796939504772804 |

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

sol: population mean = 270

n = 18

sample mean = 260

SD = 90

t = (260-270)/(90/18\*\*0.5))

= -0.471

Df = n-1 = 18-1 = 17

To find: P(X<=260)

🡪using scipy library

p\_value **=** stats**.**t**.**sf(abs(**-**0.4714) , df**=**17)

= 0.321

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 life of the bulbs is 300 days.