|  |  |
| --- | --- |
| **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 | Interval |
| Celsius Temperature | Interval |
| Weight | Ratio |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Ratio |
| Type of living accommodation | Ordinal |
| Level of Agreement | Ratio |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ratio |
| Time on a Clock with Hands | Ratio |
| Number of Children | Ratio |
| 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?**

ANS:

Three coins are tossed, All possible outcomes = [HHH, HHT, HTH, HTT, THT, THH, TTH, TTT] = 8

Outcomes with two heads and one tail = [HHT, HTH, THH] = 3

Probability of two heads and oneqas tail = Outcomes with two heads and one tail /All possible outcomes= 3/8

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

1. **Equal to 1 = 0**
2. **Less than or equal to 4 = 1/6**
3. **Sum is divisible by 2 and 3 = 1/6**

**ANS:**

All possible outcomes = {(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)} = 36

A) Outcomes of sum equal to 1 = None

Probability of sum equal to 1 = 0.

B) Outcomes of sum less than or equal to 4 = {(1,1),(1,2),(1,3),(2,1),(2,2)(3,1)}=6

Probability of sum less than or equal to 4 = Outcomes of sum less than or equal to 4/ All possible outcomes = 6/36 = 1/6.

C) Outcomes of sum is divisible by 2 and 3 = {(1,5),(2,4),(3,3),(4,2),(5,1),(6,6)}=6

Probability of sum is divisible by 2 and 3 = Outcomes of sum is divisible by 2 and 3 / All possible outcomes = 6/36 = 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:

All possible outcomes = 7C2 = 7! / (5! \* 2!) = 7\*6/2 = 21

Possible outcomes if none of balls drawn is blue = 5C2 = 5! / (3! \* 2!) = 5\*4/2 = 10

Probability that none of the balls drawn is blue= All possible outcomes/ Possible outcomes if none of balls drawn is blue = 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

Expected number of candies for a randomly selected child =∑ (x\*P(x))

= (1\*0.015) + (4\*0.2) + (3\*0.65) + (5\*0.005) + (6\*0.01) + (2\*0.12) = 3.09

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,Weight**

**Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.**

Answer:

|  |  |  |
| --- | --- | --- |
| **Points** | **Score** | **Weight** |
| Mean = 3.596563 | Mean = 3.21725 | Mean = 17.84875 |
| Median = 3.695 | Median = 3.325 | Median = 17.71 |
| Mode = 3.07 | Mode = 3.44 | Mode = 17.02 |
| Variance = 0.27694 | Variance = 0.92746 | Variance = 3.093379 |
| Standard Deviation = 0.534679 | Standard Deviation =  0.96304 | Standard Deviation =  1.7588 |
| Range = 2.17 | Range = 3.911 | Range = 8.4 |

**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: Probability of one patient chosen = P(x) = 1/9

Expected Value of the Weight of that patient= ∑(x\*P(x))

= (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)= 145.33

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

**Cars speed and distance**

**Answer:**

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | **Speed** | **Distance** | **SP** | **Weight (WT)** |
| **Skewness** | -0.11751 | 0.806895 | 1.61145 | -0.61475 |
| **Kurtosis** | -0.50899 | 0.405053 | 2.977329 | 0.950291 |

**Speed :** Negative skewness and kurtosis means data is skewed right and has wider peaks and thinner tails.

**Distance :** Positive skewness and kurtosis means data is skewed left and has thinner peaks and thicker tails.

**SP :** Positive skewness and kurtosis means data is skewed left and has thinner peaks and thicker tails.

**Weight(WT) :** Negative skewness means data is skewed right and positive kurtosis means that is has thinner peaks and thicker tails.

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





Histogram : Histogram is skewed left hence has negative skewness. As well as positive kurtosis.

Boxplot :

From boxplot we can tell that data is skewed left and as most values are more focused on lower end ie. Lower whisker, lower and upper quartil while distribution of data is less towards upper whisker. While there aren’t any outliers

beyond lower whisker, there are few outliers beyond upper whisker. From boxplot we can tell it is skewed left.

**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: Interval = x ± z (s / √n )

A) 94% confidence interval = 198.74 to 201.26.

Interval = 200± 1.8808(30\*√2000)

B) 96% confidence interval = 198.63 to 201.37.

Interval = 200± 2.0537(30\*√2000)

C) 98% confidence interval = 198.44 to 201.56.

Interval = 200± 2.3263(30\*√2000)

**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?**

ANS:

Data = [34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56]

X = np.mean(data)

Y = np.median(data)

Z = np.var(data)

W = np.std(data)

[x,y,z,w]

Mean = 41 Median = 40.5 Variance = 24.11

Standard deviation = 4.9103

Most of the students have marks near mean value. As mean > median, we can say that distribution is positively skewed.

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

= Distribution is symmetric ie. Zero skewness.

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

= Distribution is positively skewed ie. Skewed left.

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

= Distribution is negatively skewed ie. Skewed right.

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

= Positive kurtosis value indicates that distribution is peaked with thick tails.

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

= Negative kurtosis value indicates that distribution is flat with thin tails.

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



**What can we say about the distribution of the data?**

= From boxplot we can tell that most values are more focused on upper end ie. Upper whisker, upper and lower quartil while distribution of data is less towards lower whisker.

As for values 25% of values are below 10 whereas 75% of values are above 10.

**What is nature of skewness of the data?**

= For nature of skewness we can say that it is ‘Skewed right’.

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

= IQR of data will be 10 to 18 ie. 8  
  
  
**Q19) Comment on the below Boxplot visualizations?**



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

Ans :

Data for boxplot 1 is closely placed ie. Highly concentrated in short range as compare to data for boxplot 2. As data 1 is closely placed its IQR is much less than data 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**

ANS:

* 1. P(MPG>38)

Cars = pd.read\_csv(r'c:\Users\rohit\Downloads\Cars.csv')

All = cars['MPG'].value\_counts()

Print('All: ',All.sum())

A = cars['MPG']

A\_bool = (a > 38)

Print('MPG > 38: ',a\_bool.sum())

All: 81

MPG > 38: 33

P(MPG>38)=33/81

* 1. P(MPG<40)

Cars = pd.read\_csv(r'c:\Users\rohit\Downloads\Cars.csv')

All = cars['MPG'].value\_counts()

Print('All: ',All.sum())

A = cars['MPG']

A\_bool = (a < 40)

Print('MPG > 40: ',a\_bool.sum())

All: 81

MPG > 38: 61

P(MPG<40)=61/81

C. P (20<MPG<50)

Cars = pd.read\_csv(r'c:\Users\rohit\Downloads\Cars.csv')

All = cars['MPG'].value\_counts()

Print('All: ',All.sum())

A = cars['MPG']

A\_bool = (a>20) & (a<50)

Print('20 < MPG < 50: ',a\_bool.sum())

All: 81

20 < MPG < 50: 69

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

**Q 21) Check whether the data follows normal distribution**

1. **Check whether the MPG of Cars follows Normal Distribution**

**Dataset: Cars.csv**

Ans: cars = pd.read\_csv(r'c:\Users\rohit\Downloads\Cars.csv')

Cars['MPG'].hist()

From histogram if we drew a curve we won’t get a ‘Bell shaped curve’. Hence, we can say that MPG of Cars doesn’t follow 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**

Ans:

Adipose Tissue (AT) = (101.8940369724771, 96.54)

Waist Circumference(Waist) = (91.90183486238533, 90.8)

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

Ans: z-value = (1+ CL)/2

We can find corresponding Z-score from z-value using table.

|  |  |
| --- | --- |
| **Confidence interval** | **Z-Score** |
| 90% | 1.64 |
| 94% | 1.89 |
| 60% | 0.84 |

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

Ans: x = sample size = 25, df =25-1=24 , t-value = α/2 = (1-CL)/2.

We can find corresponding T-score from t-value using table.

|  |  |
| --- | --- |
| **Confidence interval** | **T-Score** |
| 95% | 2.064 |
| 96% | 2.172 |
| 99% | 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**

ANS:

Assume,

Y = average life of bulb >=270,

X = life of sample bulb = 260,

N = number of bulbs = 18,

Ρ = 90,

T-score = (260-270) / (90/√18)

T-score = -0.4714045

From T-score we get probability value P = 0.32167412.

So, probability that 18 randomly selected bulbs would have an average life of no more than 260 days is 0.32167412.