|  |  |
| --- | --- |
| 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 | Nominal |
| 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?**

**Ans:** Each coin has two outcomes

Total no of outcomes = 2\*2\*2 =8

Two heads and one tail outcomes = {HHT, HTH, THH} = 30

Probability = 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

**Ans:** a) Total outcomes = 0

Sum of two dice can never be 1. So probability = 0

b) Possible outcomes = {(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)}

Required probability =6/36=0.167

c) Sum is dividable by 2 and 3

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

P(E) = n(E) /n(s) = 5/36

**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:** There are 7 balls in the bag, and 2 of them are blue. Therefore, there are 7-2=5 balls that are not blue.

n(S) = 7C2 = 7x6 = 21

2x1

n(E) = 5C2 = 5x4 = 10

2x1

P(E) = n(E) / n(S) = 10/21 = 0.4761

**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(xi)** | **Probability(P(X=xi))** |
| **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: The expected value of a discrete random variable is the sum of the product of each possible value and its corresponding probability. It is denoted by E[X].

E[X] = ∑xi \* P(X = xi)

E[x] = (1\*0.015) + (4\*0.02) + (3\*0.65) + (5\*0.005) + (6\*0.01) + (2\*0.120)

E[x] = 0.015 + 0.08 + 1.95 + 0.025 + 0.06 +0.24

E[x] = 3.09

Therefore, the 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,Weigh>**

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

**Ans:**

|  |  |  |  |
| --- | --- | --- | --- |
| Statistic | Points | Score | weight |
| Mean | 3.5966 | 3.2173 | 17.8488 |
| Median | 3.6950 | 3.3250 | 17.7100 |
| Mode | 3.9200 | 2.7600 | 17.0200 |
| Variance | 0.2854 | 0.9573 | 3.1862 |
| Standard Deviation | 0.5347 | 0.9785 | 1.7870 |
| Range | 2.1700 | 3.8090 | 8.4000 |

Based on the calculated descriptive statistics, we can make the following inferences:

1. The Points column has the lowest deviation while the Weigh column has the highest deviation.
2. 75% of data in Points column has value less than 3.92
3. 75% of data in Score column has value less than 3.61
4. 75% of data in Weigh column has value less than 18.9

**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 weight= sum of (Weights x Probabilities) = 143.88

Expected weight of a randomly chosen patient =144 pounds

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

**Cars speed and distance**

**Use Q9\_a.csv**

|  |  |  |
| --- | --- | --- |
|  | **SPEED** | **DISTANCE** |
| **SKEWNESS** | -0.117510 | 0.806895 |
| **KURTOSIS** | -0.508994 | 0.405053 |

**Speed:** The skewness value for speed is -0.1175, which is close to 0. This indicates that the distribution of speed is relatively symmetrical. The kurtosis value for speed is -0.509, which is negative. This means that the distribution of speed is platykurtic, or has less of a peak than a normal distribution.

**Distance:** The skewness value for distance is 0.8069, which is positive. This indicates that the distribution of distance is slightly skewed to the right. The kurtosis value for distance is 0.4051, which is positive. This means that the distribution of distance is leptokurtic, or has a sharper peak than a normal distribution.

**SP and Weight (WT)**

|  |  |  |
| --- | --- | --- |
|  | **SP** | **WT** |
| **SKEWNESS** | 1.611450 | -0.614753 |
| **KURTOSIS** | 2.977329 | 0.950291 |

**SP:** The skewness value for SP is 1.611, which is positive. This indicates that the distribution of SP is skewed to the right. The kurtosis value for SP is 2.977, which is greater than 3. This means that the distribution of SP is platykurtic, or has a sharper peak than a normal distribution.

**WT:** The skewness value for WT is -0.615, which is negative. This indicates that the distribution of WT is skewed to the left. The kurtosis value for WT is 0.950, which is less than 3. This means that the distribution of WT is mesokurtic, or has a similar peak to a normal distribution.

**Use Q9\_b.csv**

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

**Ans:** This is right skewed Histogram and its skewness is seen to be positive.



**Ans:** The box plot is negatively skewed.

Q1, Median value (Q2), Q3, lower Limit, Upper Limit and Outliers

**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:** Standard error: SE = σ / , σ = 30, = 2000

SE= 30/2000 = 0.866 pounds

Margin error: ME = Z\* SE

Z= z-score for the desired confidence level

94% confidence level z-score value = 1.881

98% confidence level z-score value = 2.33

96% confidence level z-score value = 2.054

ME = 1.881\* 0.866 = 1.64

ME = 2.33\* 0.866 = 2.03

E = 2.054\* 0.866 = 1.79

Confidence intervals: Sample mean ME

For 94%, 200 pounds 1.64 = 198.36 to 201.64

For 98%, 200 pounds 2.03 = 197.97 to 202.03

For 96%, 200 pounds 1.79 = 198.21 to 201.79

**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:** 1) Mean: 41

Median: 40.5

Variance: 25.52941

Standard deviation: 5.052664

2) The student’s marks are skewed to the right, which means that there are more scores at the higher end of the distribution. The standard deviation is moderate, which means that there is a fair amount of variability in the score.

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

**Ans:** When the mean and median of a data set are equal, the data is said to be symmetrical, and the skewness is equal to zero.

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

**Ans:** When the mean is greater than the median, the data is said to be positively skewed or right-skewed.

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

**Ans:** When the median is greater than the mean, the data is said to be negatively skewed or left-skewed

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

**Ans:** A positive kurtosis value indicates that a data distribution has a sharper peak and heavier tails than a normal distribution.

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

**Ans:** A negative kurtosis value indicates that a data distribution has a flatter peak and lighter tails than a normal distribution.

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



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

**Ans:** The data is right-skewed, meaning that the tail of the distribution extends to the right. This is evident from the fact that the median is to the left of the mean, and the right whisker is longer than the left whisker.There is no outliers in the data.

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

**Ans:** The data is negatively skewed or Left Skewed

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

**Ans:** IQR = Q3 - Q1

IQR = 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: Boxplot 1** has a higher median and IQR than **Boxplot 2**, indicating a distribution of higher values and greater variability. And there is no outliers in both the boxplot.

**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)**

**Ans:** a) 0.3475939251582705

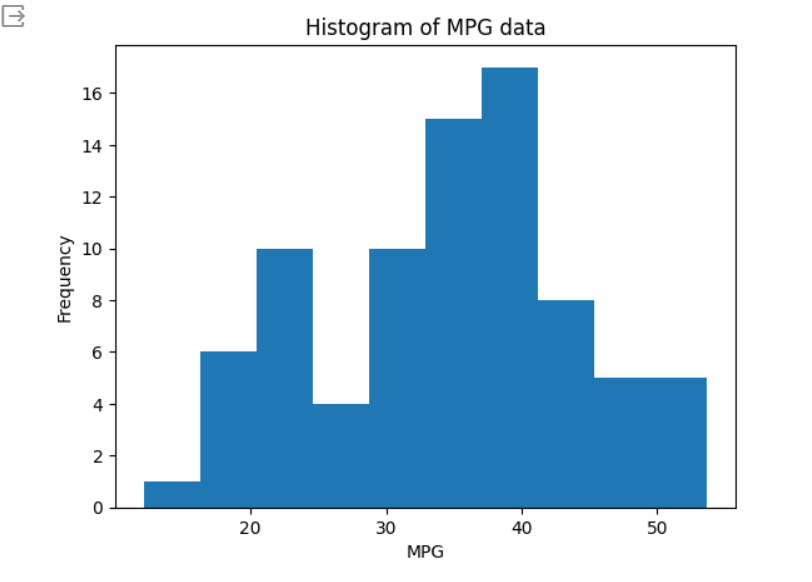
b) 0.7293498762151616

c) 0.8988689169682046

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

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

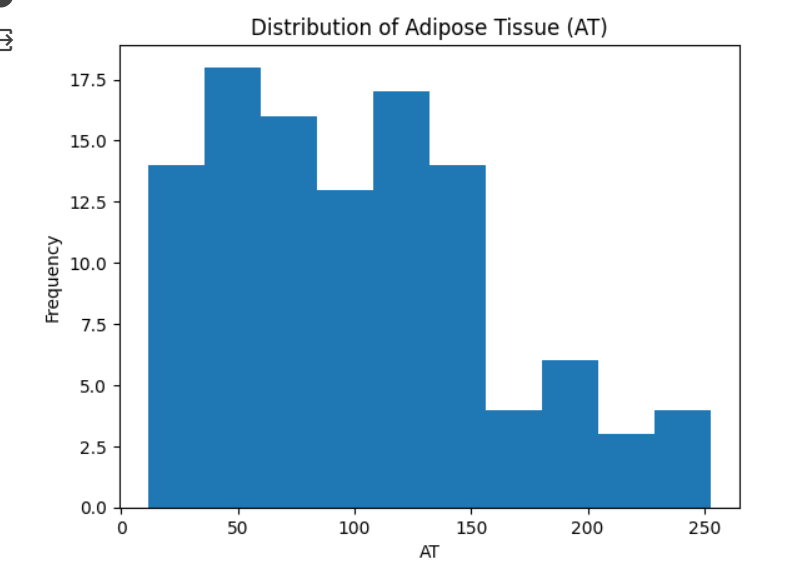
**Dataset: Cars.csv**

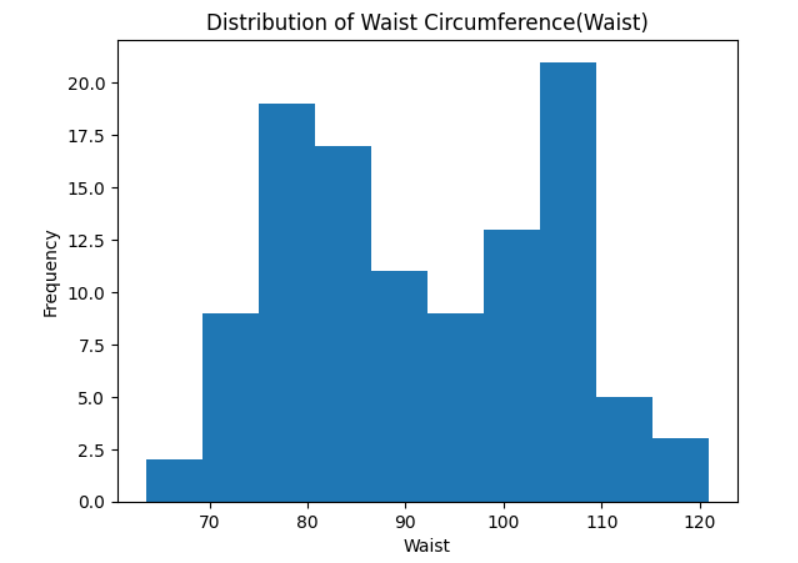
****

**Ans:** The MPG data is normally distributed. The distribution of MPG values is bell-shaped and symmetrical.

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

**Dataset: wc-at.csv**



****

**Ans:** The Adipose Tissue (AT) and Waist Circumference (WT) does not follows the Normal distribution.

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

Ans: z-score of 90% confidence interval =1.645

z-score of 94% confidence interval = 1.881

z-score of 60% confidence interval = 0.842

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

Ans: t-score of 95% confidence interval = 2.064

t-score of 96% confidence interval = 2.171

t-score of 99% confidence interval = 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**: t = -0.4714045207910317

Probability: 0.32167411684460556