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

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

**Ans:**

Total number of possible outcomes = 8

Number of favorable outcomes = 3

Probability =

= 3/8

=0.375 or 3.75%

The probability that two heads and one tail is 3.75%

**Q4) Two Dice are rolled, find is**

1. **Equal to 1 the probability that sum**

**Ans:**

The total number of outcomes when rolling two dice is = 6\*6

= 36

Total favourable cases (having sum = 1) = 0

As minimum sum is 2 for outcome (1,1).

Hence, probability is 0.

1. **Less than or equal to 4**

**Ans:**

The possible combinations that give a sum less than are equal to 4 are

(1,1),(1,2),(2,1),(3,1),(1,3) and (2,2)

The total number of outcomes when rolling two dice is = 36

Probability =

P = 6/36

= 1/6

Then the probability of Less than or equal to 4 are = 1/6

1. **Sum is divisible by 2 and 3**

**Ans:**

The combinations that result in a sum divisible by both 2 and 3 are

(1, 5), (2, 4), (3, 3), (4, 2), and (5, 1)

The total number of outcomes remains 36

Probability =

P = 5/36

Probability of the sum being divisible by 2 and 3 is 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:**

Total number of balls in a bag are =

= 21

Number of ways to draw two non blue balls =

= 10

Probability is = 10/21

= 0.4761

Then the probability that none of the balls drawn in blue is 0.4761 or 4.76%

**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**

**Ans:**

E(x) =

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

= 3.09

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**

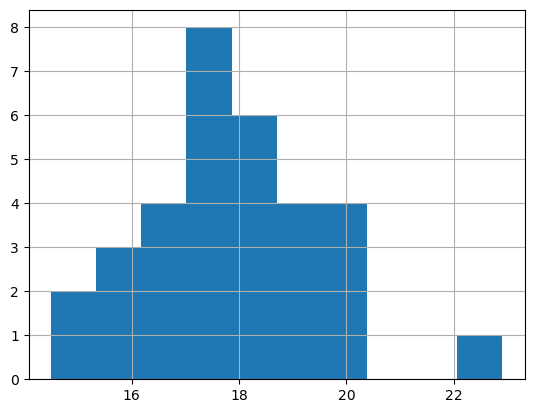
* **For Points,Score,Weigh>**

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

**Use Q7.csv file**

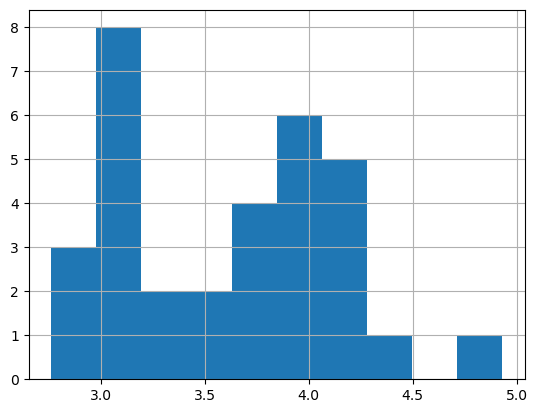
|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weight |
| Mean | 3.596 | 3.217 | 17.848 |
| Median | 3.695 | 3.325 | 17.71 |
| Mode | 3.07 | 3.25 | 17.02 |
| Variance | 0.285 | 0.957 | 3.193 |
| Standard Deviation | 0.534 | 0.978 | 1.786 |
| Range | 2.17 | 0.978 | 8.399 |
| Skew | 0.292 | 0.465 | 0.406 |

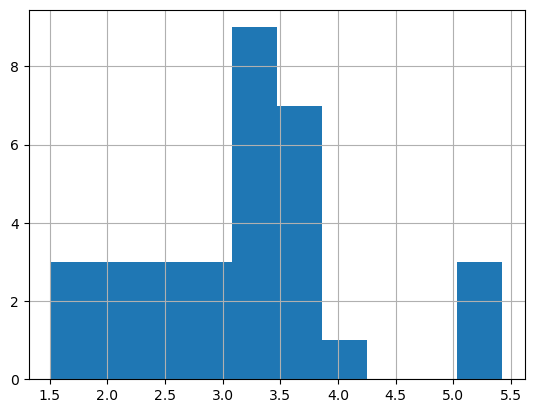
Weight histogram



Point Histogram

In [25]:



 Score histogram

**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:-**

weights = [108, 110, 123, 134, 135, 145, 167, 187, 199]

total\_sum = sum(weights)

= 108+110+123+134+135+145+167+187+199

= 1308

total\_patients = len(weights)

= 9

expected\_value = total\_sum / total\_patients

= 1308/9

= 145.333

The Expected Value of the Weight of that patient is 145.333

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

**Cars speed and distance**

**Use Q9\_a.csv**

|  |  |  |
| --- | --- | --- |
|  | Speed | Distance |
| Skew | -0.11750986 | 0.8068949 |
| Kurtosis | -0.508994420 | 0.4050585 |

**SP and Weight(WT)**

**Use Q9\_b.csv**

|  |  |  |
| --- | --- | --- |
|  | SP | WT |
| Skew | 1.61145019 | -0.6147533 |
| Kurtosis | 2.97732894 | 0.95029149 |

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



**Ans:-**

The most of the data points are concerted in the range 50-100 with frequency 200 and last range of weight is 400 somwhere around 0-10. So, the expected value the above distribution is 75. Along tail towards right, so it is heavily right skewed.



**Ans:-**

Median is less than mean right skewed and outlier on the upper side of the box plot and there is less data points between Q1 and bottom point.

**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:-**

94.0% Confidence Interval: (198.73, 201.26)

98.0% Confidence Interval: (198.43, 201.37)

96.0% Confidence Interval: (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.**

**Ans:**

Mean: 41.0

Median: 40.5

Variance: 24.11

Standard Deviation: 4.91

1. **What can we say about the student marks?**

**Ans:**

Based on these statistics, we can say that the student's marks are centered around 41.0, with some variability in the scores. The majority of scores fall within a few points of the mean, while a few scores deviate to a greater extent.

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

**Ans:-**

When the mean and median of the data are equal, the skewness is considered to be symmetrical or approximately symmetrical. This means that the data distribution is balanced

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

**Ans:-**

If the mean is greater than the median, the distribution is positively skewed.

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

**Ans:-**

When the median is greater than the mean, it suggests that the distribution is negatively skewed or left-skewed.

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

**Ans:-**

The distribution is peaked and possesses thick tails.

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

**Ans:-**

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?**

**Ans:-**

Not normally distributed

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

**Ans:**

Negative skewness

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

**Ans:**

10 - 18

**Q19) Comment on the below Boxplot visualizations?**



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

**Ans:**

By observing both the plots whisker's level is high in boxplot 2, mean and median is equal. So, distribution is symmetrical.

**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)**

**Ans:**

Probability (MPG > 38): 0.4074074074074074

* 1. **P(MPG<40)**

**Ans:**

Probability (MPG < 40): 0.7530864197530864

* 1. **P (20<MPG<50)**

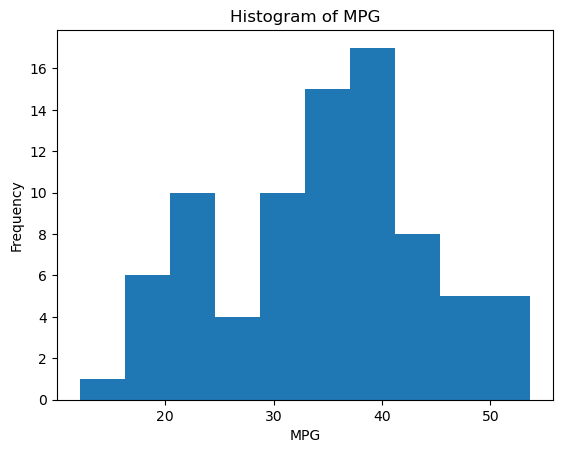
**Ans:-**

Probability (20 < MPG < 50): 0.8518518518518519

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

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

**Dataset: Cars.csv**

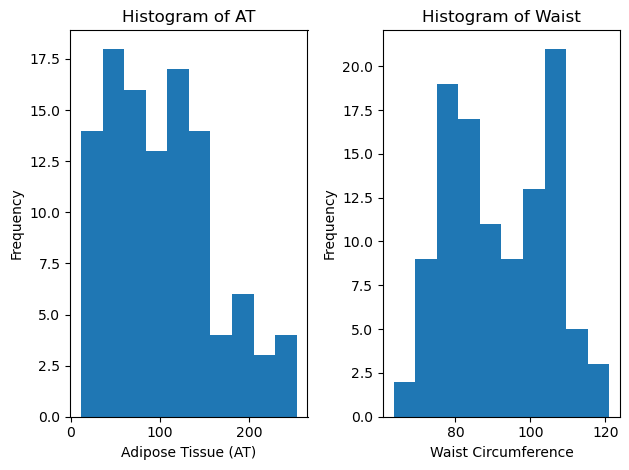


The MPG data follows a normal distribution (fail to reject H0)

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) data does not follow a normal distribution (reject H0)

The Waist Circumference data does not follow a normal distribution (reject H0)

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

**Ans:-**

90% confidence interval: Z score is ±1.645

94% confidence interval: Z score is ±1.88

60% confidence interval: Z score is ±0.84

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

**Ans:**

df=n-1= 25-1 =24

95% confidence interval, df = 24: t score is ±2.064

96% confidence interval, df = 24: t score is ±2.201

99% confidence interval, df = 24: t score is ±2.779

**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:-**

Sample size = 18 = n

Sample mean = 260 days = x

Sample standard deviation s = 90 days

= 260 – 270/90sqr(18)

= -10/9.487