|  |  |
| --- | --- |
| 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) | **Interval** |
| Sales Figures | **Interval** |
| Blood Group | **Nominal** |
| Time Of Day | **Ordinal** |
| Time on a Clock with Hands | **Interval** |
| Number of Children | **Ratio** |
| 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?

Sol: Total number of possible outcomes = 8

Number of favorable outcomes (Two heads and one tail) = 3

Probability of getting Two heads and One tail = 3/8 = **0.375**

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

Sol: The total number of possible outcomes when we roll two dice = 6\*6 = 36

Therefore, sample space = 36 = {(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)}

1. Equal to 1

Sol: **0**, as minimum sum of outcome is 2.

Let E1 be the event of probability of getting sum of equal to 1

P(E1) = n(E1)/n(s) = 0/36

**P(E1) = 0**

1. Less than or equal to 4

Sol: Let E2 be the event of probability of getting sum less than or equal to 4

S = {(1,1) (1,2) (1,3)(2,1)(2,2)(3,1)}

P(E2) = 6/36

**P(E2) = 1/6**

1. Sum is divisible by 2 and 3

Sol: Let E3 be the event of probability of getting sum divisible by 2 and 3

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

P(E3) = 6/36

**P(E3) = 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?

Sol: Total number of balls = 7

Ways of drawing 2 balls out of 7 = 7C2 = 7\*6/2\*1 = 21

Ways of drawing 2 balls (none are blue) = 5C2 = 5\*4/2\*1 = 10

**Probability of drawing 2 balls (none are blue) = 5C2/7C2 = 10/21 = 0.476**

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: Expected number of candies = E(x) = SUM {Xi\*P(Xi)}

Therefore, E(x) = 1\*0.015 + 4\*0.20 + 3\*0.65 + 5\*0.005 + 6\*0.01 = 3.090

**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.

**Use Q7.csv file**

Sol: The calculated values are provided below:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weight |
| Mean | 3.5965 | 3.2172 | 17.8487 |
| Median | 3.695 | 3.325 | 17.71 |
| Mode | 3.92 | 3.44 | 17.02 |
| Range | 2.17 | 3.911 | 8.4 |
| Variance | 0.2769 | 0.9275 | 3.0934 |
| Standard Deviation | 0.5262 | 0.9630 | 1.7588 |

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: Expected value of a random variable = E(x) = SUM {Xi\*P(Xi)}

Total number of patients = 9

Probability is every patient (since probability is not known) = 1/9

E(x) = 108/9+110/9+123/9+134/9+135/9+145/9+167/9+187/9+199/9

**E(x) = 145.33 (Expected value of weight of a random patient)**

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

**Car’s speed and distance**

**Use Q9\_a.csv**

**Sol:** The following are the skewness and kurtosis values obtained:

Skew (Speed): -0.11395477 (The Negative skew indicates that the tail is in the left side of the distribution which extends towards more negative values therefore indicating larger number of cars having high speed)

Skew (Distance): 0.78248352 (The Positive skew indicates that the tail is in the right side of the distribution which extends towards more positive values therefore indicating more cars have similar distance covered)

Kurtosis (Speed): -0.57714742 (The negative value of the kurtosis suggests that the distribution is Platykurtic)

Kurtosis (Distance): 0.24801866(The positive value of the kurtosis suggests that the distribution is Leptokurtic)

**SP and Weight (WT)**

**Use Q9\_b.csv**

**Sol:** The following are the skewness and kurtosis values obtained:

Skew (SP): 1.58145368 (The larger positive value of the skew indicates that the SP of larger number of cars are less comparatively)

Skew (Weight): -0.60330993 (The negative value of the skew indicates that the weight of cars is usually high)

Kurtosis (SP): 2.72352149 (The value of kurtosis indicates that the distribution is Mesokurtic (slightly leptokurtic))

Kurtosis (Weight): 0.81946588 (The value of kurtosis indicates that the distribution is Leptokurtic)

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



Sol: The following inferences can be drawn:

1. The distribution is positively skewed or right skewed as the tail lies towards the right of distribution.
2. The large concentration of the chicks has a weight from 50 to 100 considering the whole distribution.
3. More than 50% of the chick weight lies between 50 to 150.



Sol: The data is right skewed as the upper quartile lies closer to the bottom. The outliers can be seen at the upper side.

**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: Given data:

Average weight (Mean) = 200

Sample size (n) = 2000

Standard Deviation = 30

Range estimate = X̅ ± ∆

For confidence intervals, c=**0.94** α=1-c = 0.06 α/2= 0.03

∆ = ± Zα/2. σpop/ ∆ = ±1.26

**Range estimate = (198.73,201.26)**

Similarly, for confidence interval= **0.98**, **Range estimate = (198.43,201.57)**

Similarly, for confidence interval= **0.96**, **Range estimate = (198.61,201.39)**

**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: For the given data,

Mean = 41

Median = 40.5

Variance = 24.11 = 24 (roundoff)

Standard deviation = 4.91

Mean marks is 41. We can say that as variance is less (24) the spread of the data is less. As the standard deviation is less (4.91) the marks are more spread around the mean value.

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

Sol: A deviation from the symmetrical bell curve, or normal distribution, in a set of data is referred to as skewness. The curve is said to be skewed if it is displaced to the left or right. Skewness can be measured as an indicator of how much a distribution deviates from the normal distribution. The skew of a normal distribution is zero. A normal distribution is one in which the mean, median, and mode are equal. Therefore, when the data's mean and median are equal, the nature of skewness is zero.

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

Sol: A distortion from the symmetrical bell curve, or normal distribution, in a set of data is referred to as skewness. The curve is said to be skewed if it is displaced to the left or right. The distribution with the tail on its right side is said to be favorably skewed. For a positively skewed distribution, the value of skewness is greater than zero. Positive skewness is defined as the value of the mean being higher than both the median and the mode. Therefore, skewness has a positive aspect when mean exceeds median.

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

Sol: A deviation from the symmetrical bell curve, or normal distribution, in a set of data is referred to as skewness. The curve is said to be skewed if it is displaced to the left or right. The distribution with the tail on its left side is said to be negatively skewed. For a negatively skewed distribution, the value of skewness is less than zero. It is referred to as negative skewed when the mean value is lower than the median and mode. Therefore, skewness has a negative aspect when median > mean.

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

Sol: Kurtosis refers to the degree of presence of outliers in the distribution. It is a statistical measure, whether the data is heavy-tailed or light-tailed in a normal distribution. The excess kurtosis is used in statistics and probability theory to compare the kurtosis coefficient with that normal distribution. It can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near to zero (Mesokurtic distribution). Since normal distributions have a kurtosis of 3, excess kurtosis is calculating by subtracting kurtosis by 3. Excess kurtosis = Kurt – 3 Leptokurtic is having very long and skinny tails, which means there are more chances of outliers. Positive values of kurtosis indicate that distribution is peaked and possesses thick tails. An extreme positive kurtosis indicates a distribution where more of the numbers are located in the tails of the distribution instead of around the mean.

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

Sol: Kurtosis refers to the degree of presence of outliers in the distribution. It is a statistical measure, whether the data is heavy-tailed or light-tailed in a normal distribution. The excess kurtosis is used in statistics and probability theory to compare the kurtosis coefficient with that normal distribution. It can be positive (Leptokurtic distribution), negative (Platykurtic distribution), or near to zero (Mesokurtic distribution). Since normal distributions have a kurtosis of 3, excess kurtosis is calculating by subtracting kurtosis by 3. Excess kurtosis = Kurt – 3 Platykurtic having a lower tail and stretched around center tails means most of the data points are present in high proximity with mean. A platykurtic distribution is flatter (less peaked) when compared with the normal distribution. A distribution with a negative kurtosis value indicates that 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?

What is nature of skewness of the data?

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

1. The distribution of the data is largely spread as the minimum and maximum values cannot be seen hence can be considered as greater spread of data.
2. The upper quartile is smaller than the lower quartile, hence median is located towards the top of the data (Median>mean). It can be called as Negative skewed or left skewed.
3. The IQR of the data = 8.2 (approximately)

Q19) Comment on the below Boxplot visualizations?



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

Sol: Comparing the Boxplots 1 and 2 we can get the following inferences:

1. Boxplot 1 has a range of 50 compared to Boxplot 2 which have a range of 150.
2. Boxplot 1 has an IQR of 25 and Boxplot 2 has an IQR of 90.
3. Boxplot 1 and Boxplot 2 have the same median i.e they have the same average wbs value=263(approx.)
4. Both the boxplots have no skews i.e they are equally symmetric with respect their values.
5. Boxplot 1 has less range hence small spread of data comparatively
6. boxplot 2 has larger range hence the larger spread of data. The Boxplot 2 also has a larger Interquartile range and therefore has a greater spread in the middle 50% of the 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)

Sol: Probability (MPG>38) = 0.3476

{Code: 1-stats.norm.cdf(38,loc = cars\_df.MPG.mean(),scale = cars\_df.MPG.std()) }

* 1. P(MPG<40)

Sol: Probability (MPG<40) = 0.7293

{Code: stats.norm.cdf(40,loc = cars\_df.MPG.mean(),scale = cars\_df.MPG.std()) }

* 1. P (20<MPG<50)

Sol: Probability (20<MPG<50) = 0.8989

{Code: stats.norm.cdf(50,loc = cars\_df.MPG.mean(),scale = cars\_df.MPG.std()) -stats.norm.cdf(20,loc = cars\_df.MPG.mean(),scale = cars\_df.MPG.std()) }

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Sol: As per the given dataset, the calculated mean and median values are equal (almost equal and skewness close to zero). This concludes that the MPG of cars follow Normal Distribution.

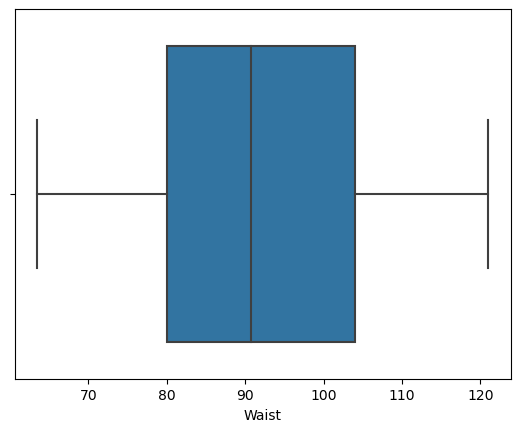
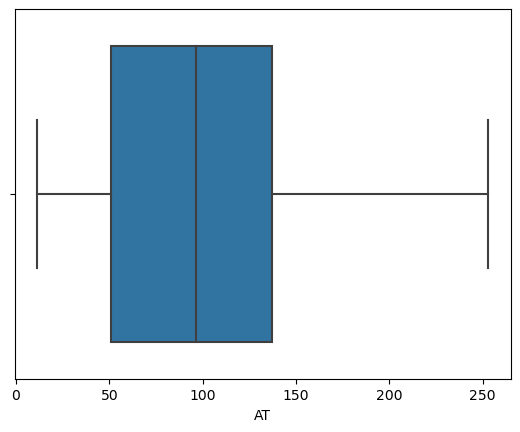
{calculated mean = 34.42, calculated median = 35.15}

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

Dataset: wc-at.csv

Sol: The Waist follows normal distribution (as mean and median almost equal and also boxplot is almost symmetrical as per the Waist boxplot)

The AT doesn’t follow a normal distribution (as per the AT boxplot below is positively skewed/ right skewed)

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

Sol: To calculate the z-scores of the given confidence intervals

1. 90% confidence interval: confidence level (c) = 0.90

Significance level (α) = 1-c = 0.10

α/2 = 0.05 Therefore Z α/2 = Z 0.05 = + 1.645

1. 94% confidence interval: c = 0.94

Significance level (α) = 0.06

α/2 = 0.03 Therefore Z α/2 = Z 0.03 = + 1.881

1. 94% confidence interval: c = 0.60

Significance level (α) = 0.40

α/2 = 0.20 Therefore Z α/2 = Z 0.20 = + 0.842

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

Sol: To calculate the t-scores of the given confidence intervals

Sample size=25, degrees of freedom = n-1 = 24

1. 95% confidence interval: confidence level (c) = 0.95

Significance level (α) = 1-0.95 = 0.05

t α/2 =t0.05/2 = t0.025 = 2.064

1. 96% confidence interval: c = 0.96

Significance level (α) = 1-0.96 = 0.04

t α/2 =t0.04/2 = t0.02 = 2.171

1. 99% confidence interval: c = 0.99

Significance level (α) = 1-0.99 = 0.01

t α/2 =t0.01/2 = t0.005 = 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

Sol: As per given data, Assume

Null Hypothesis = Ho = Average life of bulb > 260 days

Alternate Hypothesis = Ha = Average life of bulb < 260 days

Finding t-score at x=260 = t = (sample mean – population mean)

(Sample standard deviation / sqrt(n))

t = (260-270)

(90/sqrt (18))

t = -0.4714

Probability(x<260) = p\_value {taking degrees of freedom = n-1 = 17}

Using sf function, p\_value = **0.3216**