|  |  |
| --- | --- |
| **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** | Ordinal |
| **Level of Agreement** | Ordinal |
| **IQ(Intelligence Scale)** | Ratio |
| **Sales Figures** | Interval |
| **Blood Group** | Nominal |
| **Time Of Day** | Ratio |
| **Time on a Clock with Hands** | Ratio |
| **Number of Children** | Ordinal |
| **Religious Preference** | Nominal |
| **Barometer Pressure** | Ratio |
| **SAT Scores** | Ratio |
| **Years of Education** | Interval |

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

Answer: Total Events : 8 , Number of desired events : 3

P = 3/8

**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**

Answer : a) When two dice are rolled, the probability of obtaining a sum of 1 is 0, as there are no favorable outcomes out of the 36 possible outcomes.

P(sum = 1) = 0/36 = 0

b) The probability of obtaining a sum of 4 when two dice are rolled is 1/12 or approximately 0.08, with 3 favorable outcomes out of the 36 possible outcomes.

P(sum = 4) = 3/36 = 1/12 ≈ 0.08

c) When rolling two dice, the probability of obtaining a sum divisible by both 2 and 3 is 1/6 or approximately 0.16, with 6 favorable outcomes out of the 36 possible outcomes.

P(sum divisible by 2 and 3) = 6/36 = 1/6 ≈ 0.16

**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?**

Answer : The probability of drawing two balls without any of them being blue from a collection of seven balls (2 blue and 5 non-blue) can be expressed mathematically as:

P(E) = n(E) / n(S),

where P(E) is the probability of event E, n(E) is the number of favorable outcomes (10), and n(S) is the total number of possible outcomes (21). Therefore, the probability is given by:

P(E) = 10 / 21 ≈ 0.47 or 47%.

**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**

Answer : The expected number of candies for a randomly selected child, denoted as E(X), can be calculated by taking the sum of the products of each possible number of candies (x) and their corresponding probabilities (P(X=x)):

E(X) = (1 \* 0.015) + (4 \* 0.20) + (3 \* 0.65) + (95 \* 0.005) + (6 \* 0.01) + (2 \* 0.120)

This simplifies to:

E(X) = 0.015 + 0.80 + 0.195 + 0.0475 + 0.06 + 0.24

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,Weight>**

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

**Use Q7.csv file**

Answer : The data shows non-normal distribution with unequal mean, median, and mode, indicating the presence of outliers.

Mean: Points = 3.59, Score = 3.21, Weight = 17.84

Median: Points = 3.69, Score = 3.32, Weight = 17.71

Mode: Points = 3.07, Score = 3.44, Weight = 17.02

Variance: Points = 0.28, Score = 0.95, Weight = 3.19

Standard Deviation: Points = 0.53, Score = 0.97, Weight = 1.78

Range: Points = 3.59 - 4.93, Score = 3.21 - 5.42, Weight = 17.84 - 22.9

Inference: The data does not follow a normal distribution, as observed from the inequality between mean, median, and mode, and the presence of outliers.

**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?**

Answer : The expected value (E) can be calculated by taking the sum of each value (x) multiplied by its corresponding probability (P(x)):

E = P(108) \* 108 + P(110) \* 110 + P(123) \* 123 + P(134) \* 134 + P(135) \* 135 + P(145) \* 145 + P(167) \* 167 + P(187) \* 187 + P(199) \* 199

Given that the probability of selecting each patient is 1/9, the calculation simplifies to:

E = (1/9) \* (108 + 110 + 123 + 134 + 135 + 145 + 167 + 187 + 199)

Therefore, the expected value is approximately 145.33.

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

**Cars speed and distance**

**Use Q9\_a.csv**

Answer :

|  |  |  |
| --- | --- | --- |
|  | skewness | Kurtosis |
| Speed | **-**0.11 | 2.42 |
| Dist | 0.76 | 3.25 |

"dist" exhibits a positive skewness, indicating a concentration of data towards the left, while "speed" demonstrates a negative skewness, indicating a concentration of data towards the right. Additionally, both variables have positive kurtosis.

**SP and Weight(WT)**

**Use Q9\_b.csv**

|  |  |  |
| --- | --- | --- |
|  | skewness | Kurtosis |
| SP | 1.55 | 5.72 |
| WT | -0.59 | 3.82 |

"SP" displays positive skewness, implying that the data is concentrated towards the left, while "WT" exhibits negative skewness, indicating a concentration of data towards the right. This is evident from the graph. Moreover, both "WT" and "SP" have positive kurtosis.

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



Answer : The histograms show a right-skewed distribution with the peak of the data located towards the left and the tail extending towards the right. This is indicated by the fact that the mean is greater than the median. Additionally, there are outliers present on the higher end of the data. The boxplot also confirms the presence of outliers, which are located on the maximum 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?**

Answer : For a sample with a mean (x̅) of 200, standard deviation (s) of 30, and sample size (n) of 2000, the confidence interval can be calculated using the formula:

x̅ ± t \* s / √n

For a 94% confidence level, the critical value (t) is 1.55. Therefore, the interval is:

x̅ - 1.55 \* s / √n = 200 - 1.55 \* 30 / √2000 = 198.73

x̅ + 1.55 \* s / √n = 200 + 1.55 \* 30 / √2000 = 201.27

For a 96% confidence level, the critical value (t) is 1.75. The interval becomes:

x̅ - 1.75 \* s / √n = 200 - 1.75 \* 30 / √2000 = 198.61

x̅ + 1.75 \* s / √n = 200 + 1.75 \* 30 / √2000 = 201.39

For a 98% confidence level, the critical value (t) is 2.05. The interval is:

x̅ - 2.05 \* s / √n = 200 - 2.05 \* 30 / √2000 = 198.43

x̅ + 2.05 \* s / √n = 200 + 2.05 \* 30 / √2000 = 201.57

The critical values (t) were obtained from the z-table, where for a 94% confidence level, α = 0.06 and z(c) = 1.55; for a 96% confidence level, α = 0.04 and z(c) = 1.75; and for a 98% confidence level, α = 0.02 and z(c) = 2.05.

**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.**

Answer : The given statistical measures for the data are:

Mean (μ) = 41

Median = 40.5

Variance (σ^2) = 25.52

Standard Deviation (σ) = 5.05

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

Answer :

There are no extreme values that significantly deviate from the rest of the data, and the distribution of the data is slightly skewed towards the right. This can be observed by comparing the mean, which is greater than the median, indicating that the values on the higher end of the dataset pull the average towards them, causing the rightward skewness.

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

Answer : The data is symmetric and has no skewness, meaning it is evenly distributed without any notable shifts or imbalances.

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

Answer : A distribution is likely to be positively skewed if the mean is greater than the median.

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

Answer : Mean less than median implies a negatively skewed distribution.

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

Answer : Positive kurtosis: More peaked curve, heavier tails, leptokurtic.

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

Answer : Negative kurtosis: Flatter, broader curve, platykurtic.

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



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

Answer : Non-normal distribution, skewed to higher values, higher median.

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

Answer : Left-skewed data, whisker range: Minimum value extends further than maximum.

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

Answer : Interquartile Range (IQR): Upper Quartile (UQ) minus Lower Quartile (LQ) equals 8.

[IQR=UQ-LQ=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.**

Answer : There are no outliers present in the data. Both box plots have a median that falls within the range of approximately 275 to 250, indicating a similar central tendency. Additionally, the distributions are approximately normal with minimal skewness observed at both the minimum and maximum whisker range

**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)**

Answer : stats.norms.cdf(38,cars.MPG.mean(),cars.MPG.std())

P(MPG>38)= 0.3475392515

The probability of MPG being greater than 38 is approximately 0.348.

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

Answer : stats.norm.cdf(40,cars.MPG.MEAN(),cars.MPG.std())

P(MPG<40)=0.729325158

The probability of a car having an MPG value less than 40 is approximately 0.729.

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

Answer : stats.norm.cdf(0.50,cars.MPG.mean(),cars.MPG.std()) –

stats.norm.cdf(0.20,cars.MPG.mean(),cars.MPG.mean(),cars.MPG.std())

P (20<MPG<50)= 0.8518519

The probability of MPG falling within the range of 20 to 50 is approximately 0.852.

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

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

**Dataset: Cars.csv**

Answer : The MPG values of cars are assumed to be normally distributed.

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

**Dataset: wc-at.csv**

Answer : The distribution of Adipose Tissue (AT) and Waist measurements does not conform to a normal distribution.

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

Answer : #z-score of 90% confidence interval

stats.norm.ppf(0.95) = 1.6448536269

#z-score of 94% confidence interval

stats.norm.ppf(0.97) = 1.8807936081

#z-score of 60% confidence interval

stats.norm.ppf(0.6) = 0.841621233

The z-score for a 90% confidence interval is approximately 1.645, for a 94% confidence interval is approximately 1.881, and for a 60% confidence interval is approximately 0.842.

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

Answer :d.f=n-1=25-1=24

t scores of 95% confidence interval

2.060

t scores of 96% confidence interval

2.167

t scores of 99% confidence interval

2.787

For a 95% confidence interval with 24 degrees of freedom, the t-scores are approximately ±2.060. For a 96% confidence interval, the t-scores are approximately ±2.167. Lastly, for a 99% confidence interval, the t-scores are approximately ±2.787.

**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**

Answer :

from scipy import stats

from scipy.stats import norm

#find t-scores at x=260; t=(s\_mean-p\_mean)/(s\_SD/sqrt(n))

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

t

-0.4714045

#find P(X>=260) for null hypothesis

#p\_values=1-stats.t.cdf(abs(t\_scores),df=n-1)

p\_value=1-stats.t.cdf(abs(-0.4714),df=17)

p\_value

0.321674116844

A t-score of approximately -0.471 is calculated using the formula (sample mean - population mean) divided by (sample standard deviation divided by the square root of the sample size). The p-value, representing the probability of observing a value equal to or greater than 260 under the null hypothesis, is approximately 0.322.