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

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

Ans-When three coins are tossed the total number of possible outcomes are 2^3=8

These combinations are HHH, HHT, HTH, THH, TTH, THT, HTT, TTT.

The number of combinations which have two heads and one tail are:

HHT, HTH, THH which makes them 3 in number.

Therefore the Probability of getting two heads and one tails in the toss of three coins simultaneously is defined as:

P(Two heads and one tail)= Number of desired outcomes

= 3/8 or 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- Total possible outcome = n(s) =6^2=36

1. Favourable outcome = n(A) = 0

Required probability= 0/36 = 0

1. Favourable outcome = n(A) = 6

Required probability= 6/36 = 1/6

1. Favourable outcome = n(A) =5

Required probability = 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 = (2+3+2) =7

Let S be the sample space.

Then, n(s) = Number of ways drawing 2 balls out of 7

=(7\*6)/(2\*1) = 21

Let E =Event of drawing 2 balls, none of which is blue.

n(E)= Number of ways of drawing 2 balls out of (2+3) balls.

=(5\*4)/(2\*1)

=10

i.e., P(E) = n(E)/n(S) =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 = 0.2

Ans- Expected number of candies for a randomly selected child

=1\*0.015 + 4\*0.20 + 3\*0.65 + 5\*0.005 + 6\*0.01 +2\*0.12

= 0.015 + 0.8 + 1.95 + 0.025 + 0.06 + 0.24

= 3.09

Therefore, Expected number of candies for a randomly selected child= 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**

For Points=) Mean-3.60

Median-3.69

Mode-3.92

Variance-0.28

Std.Deviation-0.53

Range-2.17

For Score=) Mean-3.22

Median-3.32

Mode-3.44

Variance-0.93

Std.Deviation-0.96

Range-3.91

For Weigh=) Mean-17.82

Median-17.71

Mode-17.02

Variance-3.09

Std.Deviation-1.76

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- Expected value= ∑ (probability \* value)

= ∑P(X).E(X)

Probability of selecting each patient = 1/9

Expected value = (1/9)(108+110+123+134+135+145+167+187+199)

= (1/9)(1308)

=145.33

i.e.,Expected Value of the Weight of that patient = 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

Ans-

Skewness-

Speed =) -0.1139548

Distance =) 0.7824835

Kurtosis-

Speed =) 2.422835

Distance=) 3.248019

Here, skewness is negative, the data is negatively skewed or skewed left with respect to the speed, majority of data values greater than mean. The coefficient of skewness is 0.7823 , then the graph is said to be symmetric and data is normally distributed.

From the value of kurtosis with respect to speed and distance, the distribution is too peaked.

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Skewness=)**

**SP=) 1.581454**

**WT=) -0.6033099**

**Kurtosis=)**

**SP=) 5.723521**

**WT=) 3.819466**

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



Chick weight data is right skewed or positively skewed.

More than 50% Chick Weight is between 50 to150.

Most of the chick weight is between 50 to100.



The data is right skewed.

There are outliers at 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?

Ans- Sample mean of x͞͞͞͞͞͞͞͞ = 200

Sample standard deviation of s = 30

Sample size of n =2000

Considering a confidence interval, with 200-1=199df

The critical value is t=1.8916, hence

The 94% confidence interval is (198.73, 201.27)

The 96% confidence interval is (198.61, 201.39)

The 98% confidence interval is (198.43, 201.57)

**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-25.52

Std.Deviation-5.05

1. Average marks of students is 41. Mean is that it is significantly influenced by outliers; when this happens, the median becomes a better measure of the dataset’s center. Here is the dataset is made up of even numbers of values, then the values are 40 and 41 should be averaged to find median. Thus the median of this data set is 40.5. The variance is 25.52 i.e, the mean squared deviation of a students marks dataset. Standard deviation is calculated as the square root of the variance. That means 5.05 is the standard deviation and square root value of variance 25.52.

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

Ans- The distribution is symmetric, then the mean is equal to the median, and the distribution has zero skewness. If the distribution is both symmetric and unimodal, then the mean = median = mode.

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- If the mean is less than the median, the distribution is negatively skewed.

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

Ans-Positive excess values of kurtosis (>3) indicate that a distribution is peaked and possess thick tails.

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

Ans- Negative excess values of kurtosis (<3) indicate that a distribution is flat and has thin tails.

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



What can we say about the distribution of the data?

Ans-The distribution is negatively skewed because the whisker and half-box are longer on the left side of the median than on the right side.

What is nature of skewness of the data?

Ans- Skewness is a measure of the symmetry of a distribution. But here nature of skewness is asymmetry of the data. In an asymmetrical distribution a negative skew indicates that the tail on the left side is longer than on the right side (left-skewed).

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

Ans- Q1=10, Q2=15, Q3=18

IQR= 18-10

IQR= 8(Approx)

Q19) Comment on the below Boxplot visualizations?



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

Ans- In order to compare the average of wbs’s of boxplot 1 to those in boxplot 2. We will look at the following side-by-side boxplots, and supplement the graph with the descriptive statistics of each of the two distributions. The similarities between the two distributions are striking. Both distributions have the same center(median is 262.5 for boxplot 1 and boxplot 2 as well. However, the wbs’s of boxplot 2 have a much larger variability than the wbs’s of boxplot 1( Range: 150 Vs 48approx, IQR: 90 vs 25approx).

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)
  3. P(20<MPG<50)

Ans- a. P (MPG>38) =33/81

b. P (MPG<40) =67/81

C. 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

Mean- 34.42

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.cs

Mean-

Waist- 91.90

AT- 101.89

Median-

Waist-90.80

AT-96.54

Mode-

Waist is multimodal, AT is bimodal data.

Inference about the boxplot:

AT- Mean>median, right whisker is larger than left whisker, data is positively skewed.

Waist- Mean>median, both the whisker are of same length, median is slightly towards

left. Data is fairly symmetrically distributed.

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

Ans- Z-scores of 90% confidence interval- 1.64

Z-scores of 94% confidence interval- 1.88

Z-scores of 60% confidence interval- 0.84

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

Ans- t-scores of 95% confidence interval- 2.0639

t-scores of 96% confidence interval- 2.064

t-scores 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- The probability that t<-0.471 with 17 degrees of freedom assuming the population mean is true, the t-value is less than the t-value obtained with 17 degrees of freedom and a t-score of -0.471, the probability of the bulbs lasting less than 260 days on avg of 0.3218 assuming the mean life of the bulbs is 300 days.