|  |  |
| --- | --- |
| 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 | Interval |
| Hair Color | Nominal |
| Socioeconomic Status | Ordinal |
| Fahrenheit Temperature | Interval |
| Height | Interval |
| Type of living accommodation | Nominal |
| Level of Agreement | Interval |
| IQ(Intelligence Scale) | Interval |
| Sales Figures | Interval |
| Blood Group | Nominal |
| Time Of Day | Nominal |
| Time on a Clock with Hands | Interval |
| Number of Children | Ordinal |
| Religious Preference | Nominal |
| Barometer Pressure | Interval |
| SAT Scores | Interval |
| Years of Education | Ordinal |

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

ANS: If 3 Coins are tossed , the possible outcomes would be 8. Out of those 8 outcomes , the probability to get two heads and one tail would be 3 , i.e

(HHT , HTH , THH).

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: The formula for probability is No. of favorable outcomes / Total outcomes

a) If two dice are rolled , There could be no possible outcome that satisfies the probability of sum equals to 1, So the probability is 0/36.

b) The combination which gives sum less than or equal to 4 are 6 out of total 36 combinations. The combinations that gives sum less than or equal to 4 are (1,1) , (1,2) , (1,3) , (2,1) , (2,2) , (3,1) , So the probability is 6/36.

c) The combinations for the sum that is divisible by 2 and 3 are (3,3) and (6,6)

These are only 2 possible outcomes that could be divided by 2 and 3, So the possibility is 2/36.

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 outcomes for drawn of two balls is 21.

The probability of the 1st ball drawn would not be blue is 5/7 as there are 7 balls.

And the probability of the 2nd ball drawn would not be blue is 4/6 as there are now 6 balls left.

So the probability of 2 balls drawn at random and none of the balls is blue is 5/7 \* 4/6 = 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 = 0.20

ANS:

Number of candies = Count of candies (X) \* Probability p(X)

No. of candies = (1\*0.015 + 4\*0.20 + 3\*0.65 + 5\*0.005 + 6\*0.01 + 2\*0.120)

No. of candies = 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**

**ANS:**

Points:

Mean =3.596563, Median= 3.695, Mode= 3.92,

Variance= 0.2858814, Standard deviation= 0.5346787.

Score:

Mean= 3.21725, Median= 3.325, Mode= 3.44,

Variance= 0.957379, Standard deviation= 0.9784574

Weight:

 Mean= 17.84875, Median= 17.71, Mode= 17.02

 Variance= 3.193166, Standard deviation= 1.786943

Comment: Mean value for the columns ‘Point’ and ‘Score’ are very close.

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: To get the expected weight of a random patient we take the average of weight of all the patients. The average weight is going to be the expected weight of that random patient.

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

Avg weight = 145.33

The expected weight of a random patient is 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**ANS:** Skewness for speed is -0.11395

Here the speed is -ve which means that the graph is left skewed i.e the length of the left tail is longer.

Kurtosis for speed is 2.42285

Here the curve is peaked.

Skewness for distance is 0.7824

Here the value of skewness is +ve which means that the graph is right skewed i.e the length of the right tail is longer.

Kurtosis for speed is 3.2480

Here the curve is peaked. More than normal distribution.

**SP and Weight(WT)**

**Use Q9\_b.csv**

**ANS:** Skewness for SP is 1.581454

Here the value of skewness is +ve which means that the graph is right skewed i.e the length of the right tail is longer.

Kurtosis for SP is 2.9773

Here the curve is peaked, More than normal distribution as value of normally distributed data is 0.

Skewness for WT is -0.60331

Here the value of skewness is -ve which means that the graph is left skewed i.e the length of the left tail is longer.

Kurtosis for WT is 0.950291

Here the curve is slightly peaked, a little more than the normal distribution.

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



Here the majority of the Chick weights fall in a category of 50-100g and the maximum frequency is 200.The minimum weights have frequency less than or equal to 5.

The plot is Right skewed which shows that there is lesser concentration of chick weights in the 300-400gram category.



Here the Median is less than mean , data is right skewed and we have outlier on the upper extreme side of box plot and there are less no. of data points between Q1 and lower extreme 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:

1.- CI = 94% (μ0 – 1.04 < x < μ0 +1. 04)

2.- CI = 98% (μ0 – 2.05 < x < μ0 +2.05)

3.- CI = 96% (μ0 – 1.75< x < μ0 + 1.75)

Sample size n = 3000000

Sample mean x = 200

Standard deviation s = 30

From z-table values of z(c):

CI 94 % Confidential level α = 6 % α = 0.06 z(c) = 1.55

CI 98 % Confidential level α = 2 % α = 0,02 z(c) = 2.05

CI 96 % Confidential level α = 4 % α = 0,04 z(c) = 1.75

MOE = z(c) \* σ/√n

1.-MOE = 1.55\* 30 / √2000 MOE = 1.04

2.-MOE = 2.05\*30/√2000 MOE = 1.38

3.-MOE = 1.75\*30/√2000 MOE = 1.17

Then CI

1. CI = 94% (μ0 - MOE < x < μ0 - MOE) CI = (μ0 – 1.04 < x < μ0 +1.04)
2. CI = 98 % CI = (μ0 – 2.05 < x < μ0 + 2.05)
3. CI = 96 % CI = (μ0 – 1.75 < x < μ0 + 1.75)

**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

Median = 40.5

Variance = 5.05

Standard Deviation = 25.529

1. What can we say about the student marks?

The students have average performances. Many of the students have an average percentage of 65% and there are only a few students with the performance of above 90%.

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

When mean and median are equal then the skewness will be symmetrical.

Hence there would be no outliers for the data and the should be normally distributed.

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

When mean is greater than the median the data would be +ve skewed i.e it will show right skewness. Most of the data would be present on left side of the graph and outliers would be present on right side of the graph.

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

When median is greater than the mean the data would be -ve skewed i.e it will show left skewness. Most of the data would be present on right side of the graph and outliers would be present on left side of the graph.

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

ANS: The positive kurtosis shows that the data is sharp peaked.

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

ANS: The negative kurtosis does not show sharp peak in data, which means that the distribution of data is lighter towards tails.

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



What can we say about the distribution of the data?

ANS: The above boxplot shows median = 15.2

The Q1 value is 10 , The Q3 value is 18 and the IQR is 8

What is nature of skewness of the data?

ANS: The data is left skewed.

What will be the IQR of the data (approximately)?   
ANS: IQR = Q3 – Q1 IQR = 8

Q19) Comment on the below Boxplot visualizations?



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

ANS: Median lines of both of the box plots lie within and overlap between two boxes. The Short box mean their data points spread consistently over around the center values.

And Taller box shows more variable data and both boxes have no outliers.

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 of (MPG>38) = 0.4074074074074074

* 1. P(MPG<40)

ANS:  Probability of (MPG<40) = 0.7530864197530864

* 1. P (20<MPG<50)

ANS: Probability of (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

ANS: The data does not follow Normal distribution.

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 data does not follow Normal distribution.

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

ANS: The Z scores of 90% confidence interval = 1.645

The Z scores of 94% confidence interval = 1.8807

The Z scores of 60% confidence interval = 0.85

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

ANS: The T scores of 95% confidence interval = 2.064

 The T scores of 96% confidence interval = 2.085

 The 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 18 randomly selected bulbs would have an average life of no more than 260 days is 0.471