|  |  |
| --- | --- |
| 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 | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ordinal |
| Time on a Clock with Hands | Interval |
| Number of Children | Nominal |
| 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 ) P(2H,1T) = P(HHT)+P(HTH) +P(THH)=1/8+1/8+1/8= 3/8 = 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)

1. P(Sum = 1) = 0
2. P(Sum<= 4) = P(1,1) + P(1,2)+P(1,3)+P(2,1)+P(2,2)+P(3,1)

= 1/36+1/36+1/36+1/36+1/36+1/36

= 6/36=0.167

1. P(Sum div by 2&3) = P(1,5)+P(2,4)+P(3,3)+P(4,2)+P(5,1)+P(6,6)

= 1/36+1/36+1/36+1/36+1/36+1/36

= 6/36=0.167

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) P(No ball is blue) = P(R,R) +P(R,G)+P(G,G)

= 1/2 \* 1/2 + 1/2\*1/3+1/3\*1/3

= 1/4+1/6+1/9 = 0.528

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

|  |  |
| --- | --- |
| Mean(Points)= | 3.596563 |
| Median | 3.695 |
| Mode | 3.07 |
| SD | 0.534679 |
| Var | 0.285881 |
| Range | 2.17 |

|  |  |
| --- | --- |
| Mean(Weigh)= | 17.84875 |
| Median | 17.71 |
| Mode | 17.02 |
| SD | 1.786943 |
| Var | 3.193166 |
| Range | 8.4 |

|  |  |
| --- | --- |
| Mean(Score)= | 3.21725 |
| Median | 3.325 |
| Mode | 3.44 |
| SD | 0.978457 |
| Var | 0.957379 |
| Range | 3.911 |

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 = ∑(X\*P(X))

= (1/9)(108)+(1/9)(110)+(1/9)(123)+(1/9)(134)+ (1/9)(145)+(1/9)(167)+(1/9)(187)+(1/9)(199) = 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

**SP and Weight(WT)**

**Use Q9\_b.csv**

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



Around 200 Chicks have weight in the range of 50-100. Around 1-2 chicks have weight around 400.If we join the bins of the histogram, we can see that the distribution is Positively Skewed. As per the graph, there are no outliers present in the data.



Since the distance between median to the upper extreme is greater than the distance from median to the lower extreme, then the box plot is positively skewed. There are many outliers present on the upper extreme.

**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) S = 2000, P = 3,000,000, µ = 200, σ = 30

**94% CI**

**Z1-α = 1.5547**

**Lower Value = 200 – 1.5547 \*30/√2000 = 198.96**

**Upper Value = 200 + 1.5547 \* 30/√2000 = 201.04**

**94% CI lies within [198.96 – 201.04]**

**98% CI**

**Z1-α = 2.0537**

**Lower Value = 200 – 2.0537 \* 30/√2000 = 198.62**

**Upper Value =200 + 2.0537 \* 30/√2000 = 201.38**

**98% CI lies within [198.62 – 201.38]**

**96% CI**

**Z1-α = 1.7507**

**Lower Value = 200 – 1.7507 \* 30/√2000 = 198.82**

**Upper Value = 200 + 1.7507 \*30/√2000 = 201.17**

**96% CI lies within [198.82 – 201.17]**

**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 = 25.529

Standard Deviation = 5.0526

1. What can we say about the student marks?

Ans) 50% of the students have scored more than 40.5. As Mean > Median, there is a slight positive or Right Skewness associated. The measure of dispersion from the mean value is 5.0526 for the student marks.

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

Ans) Data is symmetrical and there is no skewness associated with it. We can say that the right tail and left tail would be equally spaced from mean and they will be mirror image of each other.

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

Ans) When mean> median, data is Positive or Right Skewed. We can also that right tail of the density graph will be more spread out than Left tail. The outliers will be present on the Right tail of distribution.

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

Ans) When median> mean, data is Negative or Left Skewed. We can also that left tail of the density graph will be more spread out than Right tail. The outliers will be present on the Left tail of distribution.

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

Ans) The distributions with Positive Kurtosis have heavy tails that are longer and contain more extreme values. In short, there is a greater tendency for outliers. They have values of greater than 3 or positive excess values (> 0).

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

Ans) The distribution with Negative Kurtosis have lighter tails that are shorter and contain fewer outliers. These distributions have values of less than 3 or negative excess values (< 0).

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



What can we say about the distribution of the data?

* 25% of data values lies below 10(i.e Q1), 75 % of data values lies below 18(i.e Q3) and 50 % of data values lie below around 15(i.e also the median of data).
* Lower Extreme = Q1 –(1.5\*IQR) = -2
* Upper Extreme = Q3 +(1.5\*IQR) = 30
* Outliers might be present either below -2 and above 30.

What is nature of skewness of the data?

* The vertical line inside the box that represents the median is much closer to the third quartile than the first quartile, which means the distribution is left-skewed.

What will be the IQR of the data (approximately)?   
  
 -- IQR = Q3-Q1 = 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.

Ans) Medians of both the data groups is around 257.

Here 2 has a longer length than 1 which means that the dispersion of data is more in 2 as compared to 1. Plot 2 has larger range than Plot 1. The data in both the groups is almost symmetrical with no skewness.

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)

c. P (20<MPG<50)

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

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

Dataset: wc-at.csv

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

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

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) N = 18, X = 260, σ = 90 ,µ = 270

Step 1:

H0 = Light bulb last >= 260

HA = Light bulb last < 260

Step 2: t – Test

Step 3: One-Tail

Step 4: α = 0.05 (CI 95%)

Step 5:

t = -0.4714 ,p value = 0.3217

As p value is greater than α, Accept Null hypothesis, Average life of Bulb is greater than 260 days