|  |  |
| --- | --- |
| 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 | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Ratio |
| Years of Education | Ratio |

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

Ans: S = {HHH, HHT, HTT, HTH, THT, TTT, TTH, THH}

P (two heads and one tail) = 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: all possible outcomes = 6^2 = 36

a) P (sum is equal to 1) = 0 …..(Since minimum sum would be 1+1= 2)

b) All possible outcomes for sum less than or equal to 4 = {(1,1), (1,2), (1,3), (2,1), (2,2), (3,1)}

P ( sum less than or equal to 4) = 6/36 = 0.1667

c) Sum is divisible by 2 and 3 = {(1,5), (2,4), (3,3), (4,2), (5,1), (6,6)}

P(Sum is divisible by 2 and 3) = 6/36 = 0.1667

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: n(S) = 7

P(none of the balls drawn is blue) = 1 - P(atleast one ball drawn is blue)

= 1 – 1/7

= 0.857

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: Expected number of candies for a randomly selected child

= (1 x 0.015) + (4 x 0.20) + (3 x 0.65) + (5 x 0.005) + (6 x 0.01) + (2 x 0.12)

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

= 3.09

= 3 candies

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.

Ans:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| mean | 3.596563 | 3.217250 | 17.848750 |
| median | 3.6950 | 3.325 | 17.71 |
| mode | 3.07 | 3.44 | 17.02 |
| variance | 0.2769 | 0.92746 | 3.0933 |
| standard deviation | 0.534679 | 0.978457 | 1.786943 |
| range | 2.17 | 3.91100 | 8.3999 |

In “Points” data is median > mean hence we can say it is sightly right skewed as difference is not much, this implies that there may be an outlier present on the higher side of values. As standard deviation is small so is the range hence the spread of data is lower.

In “Score” data is median > mean hence we can say it is sightly right skewed as difference is not much, this implies that there may be an outlier present on the higher side of values.

In “Weigh” data is median < mean hence we can say it is sightly left skewed as difference is not much, this implies that there may be an outlier present on the lower side of values.

Also, the there is moderate negative correlation between points and score, so can roughly say that they are inversely proportional.

**Use Q7.csv file**

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 of the Weight of that patient = 145.333

**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**

Ans:

|  |  |  |
| --- | --- | --- |
| Q9.a | Skewness | Kurtosis |
| speed | -0.11395 | -0.5771 |
| distance | 0.78248 | 0.24801 |
| “speed” the skewness is between -0.5 & 0.5, the data are nearly symmetrical which means that mean < median, maybe a few outliers on the lower side of the data points. The kurtosis < 3 which means it is Platykurtic having a lower tail and stretched around center tails means most of the data points are present in high proximity with mean.  Data in “distance” is between 0.5 & 1(positive skewed), the data are slightly skewed means that mean > median due to moderate amount of data is populated on the higher side of the values and some outliers on the higher side of the values. The kurtosis < 3 which means it is Platykurtic having a lower tail and stretched around center tails means most of the data points are present in high proximity with mean. | | |
| Q9.b | | |
| SP | 1.5814 | 2.7235 |
| WT | -0.6033 | 0.81946 |
| “SP” data is extremely positively skewed which means that mean > median, a large number of data-pushed on the right-hand side and some extreme outliers on the higher side of the values. The kurtosis < 3 which means it is Platykurtic having a lower tail and stretched around center tails means most of the data points are present in high proximity with mean.  Data in “WT” is sightly negatively skewed means that mean > median due to moderate amount of data is populated on the lower side of the values and some outliers on the lower side of the values The kurtosis < 3 which means it is Platykurtic having a lower tail and stretched around center tails means most of the data points are present in high proximity with mean. | | |

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

 Ans:

* The histogram looks pulled to the right as the distribution is right skewed.
* Majority of chicken population seem to weigh between 0 to 200, most of this population being in range of 50 to 150.
* A minor population of about less than 50 chickens are above 300 in weight.



Ans: The distance between the 1st quantile and median is less than 3rd quantile and median, also there are some huge outliers on the upper side suggesting that the distribution is right skewed.

**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: 94% Confidence internal = (198.7376089443071, 201.2623910556929)

98% Confidence internal = (198.4381860483216, 201.5618139516784)

96% Confidence internal = (198.6214037429732, 201.3785962570268)

**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: mean = 41

median = 40.5

variance = 24.111

standard deviation = 5.05267

Minimum marks scored were 34 and maximum marks scored were 56.

50% of marks scored were between 38 and 42 marks.

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

Ans: The distribution is normal and symmetric, there is no skewedness.

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

Ans: Right skewed

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

Ans: Left skewed

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

Ans: It indicates that the data has high variance and some extreme outliers.

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

Ans: It indicates that the data has low variance and there may be some outliers, which are not that extreme.

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



What can we say about the distribution of the data?

Ans: It is a normally distributed data with high variance and outliers on lower side of the values, mean < median.

What is nature of skewness of the data?

Ans: It is left skewed.

What will be the IQR of the data (approximately)?   
Ans: IQR = 75th quantile – 25th quantile

=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: The data of box plot 1 has low variance as compared to the 2nd box plot as the range of box plot 1 is contained within the inter-quantile range of box plot 2. Both the box plots are symmetrical in nature suggesting normal distribution of 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) = 33/81

= 0.4074

* 1. P(MPG<40) = 61/81

= 0.7531

c. P (20<MPG<50) = 69/81

= 0.852

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

(Solved in Juypter notebook)

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

Dataset: wc-at.csv

(Solved in Juypter notebook)

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

Ans:

# 90% CONFIDENCE INTERVAL = 1.6448536269514722

# 96% CONFIDENCE INTERVAL = 1.8807936081512509

# 60% CONFIDENCE INTERVAL = 0.8416212335729143

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

Ans:

# 90% CONFIDENCE INTERVAL = (-2.063898561, 2.063898561)

# 96% CONFIDENCE INTERVAL = (-2.3069134, 2.3069134)

# 60% CONFIDENCE INTERVAL = (-1.31783593, 1.31783593)

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 à pt(tscore,df)

df à degrees of freedom

Ans: (Solved in Juypter notebook)