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

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

probability = no of instances for 2 heads & 1 tail / total no of instances = 3/8

**Ans: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

Ans: a.) Probability = Expected instances/total instances = 0/36 = 0

b.) Probability = Expected instances/total instances = 6/36 = 1/6

c.) Probability = Expected instances/total instances = 6/36 = 1/6

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: Probability = Expected instances/total instances = (5C2)/(7C2) = 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

Expected number of candies: Ʃ(P\*X) = Ʃ(1\*0.015 + 4\*0.20 + 3\*0.65 + 5\*0.005 + 6\*0.01 + 2\*0.120) = 3.090

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**

**Mean**

Points 3.596563

Score 3.217250

Weigh 17.848750

**--------------------**

**Median**

Points 3.695

Score 3.325

Weigh 17.710

**-------------------**

**Mode**

Points – (3.07, 3.92)

Score – 3.44

Weigh – (17.02, 18.90)

--------------------

**Range**

The range of POINTS is 2.76 to 4.93

The range of SCORE is 1.513 to 5.424

The range of WEIGH is 14.5 to 22.9

`

------------------------

**Variance and std deviation**

The variance and std dev of POINTS are 0.2858813508064516 and 0.5346787360709715 resp.

The variance and std dev of SCORE are 0.9573789677419354 and 0.9784574429896966 resp.

The variance and std dev of WEIGH are 3.193166129032258 and 1.7869432360968431 resp.

**COMMENTS:**

The mean and median Points are same. Also, mode of Score is a single value means both mode values of Score are the same. The Score is more dispersed as compared to other two

columns.

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: **145.33**

a = [108, 110, 123, 134, 135, 145, 167, 187, 199]

sum = 0

for i in a:

sum+= ((1/9)\*i)

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

**Cars speed and distance**

**Use Q9\_a.csv**

Skewness of Speed = -0.1139

Skewness of distance = 0.7824(slightly skewed)

The speed data is nearly symmetrical whereas distance data is slightly positively skewed.

--------------------------

Kurtosis of Speed = -0.5771

Kurtosis of distance = 0.2480

The kurtosis of speed has lighter tails than the normal distribution. The  most data points are present in high proximity to the mean in both ways.

**SP and Weight(WT)**

**Use Q9\_b.csv**

Skewness of SP = 1.5814

Skewness of WT = -0.6033

The SP data is highly skewed as right side whereas WT data is slightly negatively skewed.

--------------------------

Kurtosis of SP = 2.72

Kurtosis of WT = 0.819

The kurtosis of SP data has higher tail than WT data. The most WT data points are present in high proximity to the mean.

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



The frequency is high on weight range 50-100 and then drops silently after 100.

This is positively skewed, or right-skewed distribution has a long right tail. It is a sort of distribution where the measures are dispersing, unlike symmetrically distributed data where all measures of the central tendency (mean, median, and mode) equal each other. This makes Positively Skewed Distribution a type of distribution where the mean, median, and mode of the distribution are positive rather than negative or zero.



The above box plot has some suspected outliers lie in the frame. The outliers are observed above maximum values of box plot. The median is slightly closer to the lower quartile Q1 as compared to Q3.

**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?

**(143.57619175546247,256.42380824453755)94%**

**(130.2095637787748, 269.7904362212252)98%**

**(138.38753268104531, 261.61246731895466)96%**

**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.

Mean-41

Median-40.5

Mode- 41

Variance- 25.52

Sd – 5.05

1. What can we say about the student marks?

maximum students got 41 marks. The average is 41 marks and data is nearly centered to 41.

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

* It will be perfectly symmetrical graph with no skewness

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

* We can have then right skewed data.

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

* We can see then data skewed to the left.

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

* Positive values of kurtosis indicate that distribution is peaked and possesses thick tail. More values of the datapoints are located at the tails of the distribution rather than on its mean and outliers are present.

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

* negative values of kurtosis indicate that distribution is not normal and possesses thin tail. More values of data points are located at the mean of the sample.

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



What can we say about the distribution of the data?

* The given data is probably left skewed. 50% of the datapoints lies between 10 to 18.

What is nature of skewness of the data?

* As per the plot, the data is left side skewed.

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

* IQR is between 10 to 18

Q19) Comment on the below Boxplot visualizations?



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

* The IQR size is more in 2nd boxplot compared to the first. But Also at the same time the median is equal for both the plots which is nearly 263. Also we can conclude that both are nearly symmetrically(normally) distributed.

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) - **0.3475**
  2. P(MPG<40) – **0.7293**

c. P (20<MPG<50) – **1.24e-05**

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

* No. It is not symmetric on both sides of mean. Also median is greater than the mean means is has right skew.

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

Dataset: wc-at.csv

* Do not follow as the mean and median are not same

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

* **1.644**
* **1.880**
* **0.841**

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

* **2.0638**
* **2.1715**
* **2.7969**

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

* **0.3216**