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

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

A:3/8 or 0.375

Q4) Two Dice are rolled, find the probability that sum is

1. Equal to 1 = 0
2. Less than or equal to 4 = 1/6 or 0.167
3. Sum is divisible by 2 and 3 = 1/6 or 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?

A: 5/2\*4/6 = 20/42 = 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

A: Excepted number = 1 \* 0.015 + 4 \* 0.20 + 3 \* 0.65 + 5 \* 0.005 + 6 \* 0.01 + 2 \* 0.120 = 0.015 + 0.80 + 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.

A:

For Points :

Mean = 3.597

Median = 3.695

Mode = 3.92

Standard Deviation = 0.5346787

Variance = 0.2858814

Range = 4.930 - 2.760 = 2.17

The difference between mean and median is less, so the data is considered to be normal data and distribution is symmetric.

Low Standard Deviation and Low Variance specifies the most of the data points lies surrounding the mean.

For Score :

Mean = 3.217

Median = 3.325

Mode = 3.44

Standard Deviation = 0.9784574

Variance = 0.957379

Range = 5.424 - 1.513 = 3.193

The difference between mean and median is less, so the data is considered to be normal data and distribution is symmetric.

Low Standard Deviation and Low Variance specifies the most of the data points lies surrounding the mean.

For Weigh :

Mean = 17.85

Median = 17.71

Mode = 17.02

Standard Deviation = 1.786943

Variance = 3.193166

Range = 22.90 - 14.50 =8.399

The difference between mean and median is less, so the data is considered to be normal data and distribution is symmetric.

**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?

A: Given that

No. of patients = 9

Probability of each patient drawn one at random = 1/9

Expected Value = Summation P(x) \* E(x)

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

=1/9\* 1308

=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**

**A:**

Skewness of Car speed and Distance = -0.11750, 0.80689

Kurtosis of Car speed and Distance = -0.50899, 0.40505

Skewness of Car speed is negatively or left skewed whereas Skewness of Car distance is positively or right skewed.

The value of Skewness of Car speed is between -0.5 and 0.5, the data are fairly symmetrical.

The value of Skewness of Car Distance is between 0.5 and 1, the data are moderately skewed

Kurtosis of Car speed- the distribution is broad and lighter tails on either side are produced.

Kurtosis of Car distance- the distribution is narrow and higher tails on either side are produced.

Skewness of SP and WT = 1.6114, -0.6147

The value of Skewness of SP is greater than 1, the data are highly skewed.

The value of Skewness of WT is between -1 and – 0.5, the data are moderately skewed.

Kurtosis of SP and WT = 2.9773, 0.9502

Skewness of SP is positively or right skewed whereas Skewness of WT is negatively or left skewed.

Kurtosis of SP and WT - the distribution is narrow and lighter tails on either side are produced.

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

**A:** The below Histogram graph, the data is distributed more towards the right side. So, the Data is right skewed with implies to be positively skewed.

The highest frequency of around 198 is present at the interval 50 – 100.

Box plot depicts outliers above the upper fence.



**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?

A: Given data –

n= 2000

X(bar) = 200

Std deviation = 30

Confidence Interval for 94% - 200 + 1.88 \* 0.6708 = 201.2611

200 - 1.88 \* 0.6708 = 198.7389

Confidence Interval for 98% - 200 + 2.33 \* 0.6708 = 201.5629

200 - 2.33 \* 0.6708 = 198.4371

Confidence Interval for 96% - 200 + 2.05 \* 0.6708 = 201.3751

200 - 2.05 \* 0.6708 = 198.6248

**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.

A: Mean - 41

Median - 40.5

Variance – 25.52

Standard Deviation – 5.052

1. What can we say about the student marks?

A: The difference between mean and median is less. So, the data is considered to be normal data and the is distribution is symmetric around the mean.

Outliers can be present on the right as the last values i. e. marks are slight deviated considering whole dataset.

As Standard Deviation and Variance are low then, data specifies the most of

the datapoints lies surrounding the mean.

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

A: When mean and median are equal then, data distribution represents Zero

Skewness.

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

A: When mean is greater than median then, data distribution represents

Positive Skewness or Right Skewness.

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

A: When median is greater than mean then, data distribution represents

Negative Skewness or Left Skewness.

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

A: It indicates the graph with narrow peak and broader tails

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

A: It indicates the graph with wide peak and thinner tails

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



What can we say about the distribution of the data?

A: The data is has non symmetrical distribution. The data is not equally distributed between the Quartiles and the lengths of whiskers also differ from each other.

What is nature of skewness of the data?

A: The data distribution is left skewed which is termed as negatively skewness.

Therefore, Outliers might be present at the tail on the left side.

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

A: The Inter Quartile Range = 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.

A: The boxplot 2 is normally distributed as compared to boxplot 1 as the quartiles in boxplot 2 are nearly equal whereas the upper quartile is slightly large than the lower quartile.

The median value for both the box plots is nearly equal to 262.

There are no outliers depicted in both the boxplots.

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)

A: In python-

1-stats.norm.cdf(38,

loc=df.MPG.mean(),

scale=df.MPG.std())

=0.3475

* 1. P(MPG<40)

A: In python-

stats.norm.cdf(40,

loc=df.MPG.mean(),

scale=df.MPG.std())

=0.7293

* 1. P (20<MPG<50)

A: In python-

a=1-stats.norm.cdf(20,

loc=df.MPG.mean(),

scale=df.MPG.std())

b=stats.norm.cdf(50,

loc=df.MPG.mean(),

scale=df.MPG.std())

c=b-a

c

=0.0131

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

A: The data is distributed normally as the difference mean, median and mode is very less.

Skewness value is less than 1, the data are moderately skewed.

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

Dataset: wc-at.csv

A: The data is distributed normally as the difference mean, median and mode is very less.

Skewness value is less than 1, the data are moderately skewed.

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

A: Z Score of 90% confidence interval = 1.6449

Z Score of 94% confidence interval = 1.8808

Z Score of 60% confidence interval = 0.8416

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

A: T Score of 95% confidence interval = 2.0639

T Score of 96% confidence interval = 2.1715

T Score of 99% confidence interval = 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

A: Given -

Population mean = 270

n = 18

Sample mean = 260

Sample standard deviation = 90

T score = 260-270/90/sqrt(18)

=-0.471

stats.t.cdf(-0.471,17) or 1-stats.t.cdf(0.471,17)

=0.3218 ~0.322