|  |  |
| --- | --- |
| 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?

**Ans:-**p (2H, 1T) = (total no of favorable outcomes) / (Total no of outcomes) = 3/8 = 0.375.

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

Equal to 1

**Ans: -** 0

Less than or equal to 4

**Ans: -** 1/6

Sum is divisible by 2 and 3.

**Ans: -** 1/6

Q5) is there any shortcut trick of finding factorial values like 7!

**Ans: -** By using nCr formula = (n!)/((r!) (n-r)!). where n is total no of items and r is no of items to choose.

Total no of balls = (2R + 3G +2B) = 7 balls.

No of way to draw 2 balls = 7C2 = 7! / ((2!) (7-2)!) = 21.

No of ways to draw 2 balls drawn are not blue involves red and blue balls (5 balls)  
 5C2 = 5! / (2!) (5-2)!) = 10.

**P (no blue)** = 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: -** Expected no of candies = candies count of (A, B, C, D, E, F) \* Probability of (A, B, C, D, E, F) respectively, i.e., 3.12

Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset.

For Points, Score, Weight

Find Mean, Median, Mode, Variance, Standard Deviation, and Range and Comment about the values/ Draw some inferences.

Use Q7.csv file.

**Ans: -**

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | Points | Score | | Weight | |
| mean | 3.596563 | 3.217250 | | 17.848750 | |
| median | 3.695000 | 3.325000 | | 17.710000 | |
| mode | 3.07 | 3.44 | | 17.02 | |
|  | 3.92 | NaN | | 18.90 | |
| std | 0.534679 | | 0.978457 | | 1.786943 |
| Range | 4.644119 | | 4.466621 | | 21.113057 |
| var | 0.285881 | | 0.957379 | | 3.193166 |

When we compare mean, median, mode there are almost same so, the data follows normal distribution.

Q8) Calculate Expected Value for the problem below.

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: -** Calculate the sum of all patient’s weight.

Then, divide the sum with total no of patients(n).

The sum of all patients’ weight is 1308 pounds.

Total no of patients is (n) = 9.

The expected weight of the random patient chosen is near to 145.333333.

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

Car’s speed and distance

Use Q9\_a.csv  **Inferences**

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| |  |  | | --- | --- | | S/No | speed dist | | 1. count | 50.000000 50.000000 | | 1. mean | 15.400000 42.980000 | | 1. std | 5.287644 25.769377 | | 1. min | 4.000000 2.000000 | | 1. 25% | 12.000000 26.000000 | | 1. 50% | 15.000000 36.000000 | | 1. 75% | 19.000000 56.000000 | | 1. max | 25.000000 120.000000 | |

**Ans :-**

Skewness :-

Skewness of dataset\_1:

speed -0.117510

dist 0.806895

Kurtosis of dataset\_1:

speed -0.508994

dist 0.405053

SP and Weight (WT)

Use Q9\_b.csv

**Ans: -** Skewness of dataset\_2:

SP 1.611450

WT -0.614753

Kurtosis of dataset\_2:

SP 2.977329

WT 0.950291

Q10) Draw inferences about the following boxplot & histogram.



**Ans: -** The above histogram represents the distribution of chick weights. The x-axis is labeled as “ChickWeight$weight” and ranges from 0 to 400. The y-axis is labeled as “Frequency” and ranges from 0 to 200. There are seven bars representing different weight categories. The second bar, representing weights from about 50 to 100, reaches above 150 frequency. Subsequent bars decrease in height as they move right along the x-axis, indicating reduced frequencies for higher weight categories*.* A majority of chicks have weights between 50 and 150 grams, with the highest frequency occurring in the 50–100-gram range. The frequency decreases as the weight increases, indicating that may be there are fewer chicks in the higher weight categories.



**Ans: -** The bottom and top of the box represent the (Q1 and Q3) respectively, and the line inside the box is the median (Q2). The whiskers extend to show the range of the data, excluding outliers which are plotted individually as points above the upper whisker.

In this boxplot, the median line is closer to lower wisher, indicating that data is **positively skewed** and there are more than seven (approx.) **outlier points** plotted above the upper whisker, indicating that there are **some extreme values** in the dataset.

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: -** Given S\_mean(x\_bar) = 200 pounds

S\_size(n)= 2000 men, P\_size(N) = 3,000,000 men

S\_standard deviation(s) = 30 pounds

Formula: - Confidence Interval = 1-α/2, n-1 s/n0.5

Where α = Significance level

Case\_94%: - α = 100% - 94% = 6% = 0.06 ; α/2 = 0.03

Case\_98%: - α = 100% - 98% = 2% = 0.02 ; α/2 = 0.01

Case\_96%: - α = 100% - 96% = 4% = 0.04 ; α/2 = 0.02

t\_values = [1.881, 2.328, 2.055] at 94%, 98%, 96% respectively.

This t values obtained from python using “stats.t.ppf (α\_value, n-1)”  
By sub t\_values in above Formula we get confidence interval values:

94% Confidence Interval is **[198.7376089443071, 201.2623910556929]**

98% Confidence Interval is **[198.4381860483216, 201.5618139516784]**

96% Confidence Interval is **[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

Find mean, median, variance, standard deviation.

What can we say about the student marks?

**Ans: -** Mean = 41.0

Median = 40.5

Variance = 24.11

Sta\_deviation = 4.91

The mean score indicates that the student’s performance is above average. The median score suggests that the student’s performance is consistent. The variance and standard deviation indicate that the scores are spread out and not tightly clustered around the mean score. This suggests that the student’s performance is not consistent across all tests.

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

**Ans:** **-** When mean and Median are equal the nature of the skewness is Symmetric.

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

**Ans: -** When mean and Median are equal the nature of the skewness is Right skewed.

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

**Ans: -** When mean and Median are equal the nature of the skewness is Left skewed.

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

**Ans: -** Positive kurtosis value indicates that the data is distributed Laplace (continuous and high tail) and outlier frequency is high.

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

**Ans: -** Negative Kurtosis value indicates that the data is normal distribution having thin tailed and outlier frequency is low.

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



What can we say about the distribution of data?

**Ans: -** Data is Negatively skewed because median is closer to the top of the box and whisker is longer than right.

What is the nature of skewness of the data?

**Ans: -** Negative skewness.

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

**Ans: -** IQR (Q3 -Q1) = 10(approx.)  
  
  
Q19) Comment on the below Boxplot visualizations?



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

**Ans :-** **Median Values** = Both the data have the same median value.

**IQR** = Boxplot\_1 has smaller IQR, whereas Noxplot\_2 has larger IQR i.e., in boxplot\_1 the data is distributed with less variability than boxplot\_2.

**Skewness** = no skew, because median line at center of the box then the distribution is symmetrical.

**Distribution** = Evenly distributed around median value.

**Outliers** = no significant outliers on either side of the distribution.

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

P(MPG>38)

P(MPG<40)

P (20<MPG<50)

**Ans: -** Probability of MPG > 38 is: 0.4074.

Probability of MPG < 40 is: 0.7531.

Probability of 20 < MPG < 50 is: 0.8519.

Q 21) Check whether the data follows normal distribution

Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Ans: -** Given data (Cars.csv) does not follow normal distribution.

MPG of Cars dataset also does not follow normal distribution because, the histogram and curve does not follow a normal distribution properties. The data is skewed to the right side, meaning it’s not symmetrical around its mean.

Q 22) Check Whether the Adipose Tissue (AT) and Waist Circumference (Waist) from

wc-at data set follows Normal Distribution

Use Dataset: wc-at.csv

**Ans: -** From the histogram, given data does not follow Normal Distribution.

Q 23) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence interval ?

**Ans: -** The Z score for a 90% confidence interval is 1.64.

The Z score for a 94% confidence interval is 1.88.

The Z score for a 60% confidence interval is 0.84.

Q 24) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence interval for sample size of 25

**Ans: -** The T score for a 90% confidence interval is 1.71.

The T score for a 94% confidence interval is 1.97.

The T score for a 60% confidence interval is 0.86.

Q 25) 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 **approximately 0.322.**