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

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

A={(HHH),(HHT),(HTH),(THH),(HTT),(THT),(TTH),(TTT)}

P(A)= 8

B= Probability of two heads and one tail:-

B={(HHT),(HTH),(THH)}

P(B)= 3/8

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

A= {(1,1),(1,2),(1,3),(1,4),(1,5),(1,6),(2,1),(2,2),(2,3),(2,4),(2,5),(2,6),(3,1),(3,2),(3,3),(3,4),(3,5),(3,6),(4,1),(4,2),(4,3),(4,4),(4,5),(4,6),(5,1),(5,2),(5,3),(5,4),(5,5),(5,6),(6,1),(6,2),(6,3),(6,4),(6,5),(6,6)}

P(A)=36

1. Equal to 1 = 0
2. Less than or equal to 4 = B={(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)}

P(B)= 6/36 OR 1/6

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

P(C)=6/36 OR 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?

A={Red,Red,Green,Green,Green,Blue,Blue}

P(A)=7

If none of balls drawn is blue:-

B={(Red,Red,Blue,Blue),(Green,Green,Green,Blue,Blue)}

B=(4/7,5/7)

P(B)= 20/42 OR 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=

1\*0.015+4\*0.20+3\*0.65+5\*0.005+6\*0.01+2\*0.120

= 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**

**Mean of points in given data:-** 3.596563

**Median of points in given data:-** 3.695

**Mode of points in given data:-** Numeric

**Standard derivation of points in given data:-** 0.534679

**Varience of points in given data:-** 0.285881

**Range of points in given data:-** 2.17

**Mean of score in given data:-** 3.217250

**Median of score in given data:-** 3.325

**Mode of score in given data:-** Numeric

**Standard derivarion of score in given data:-** 0.978457

**Varience of score in given data:-** 0.957379

**Range of score in given data:-** 3.9110000000000005

**Mean of weight in given data:-** 17.848750

**Median of weight in given data:-** 17.710

**Mode of weight in given data:-** Numeric

**Standard derivation of weight in given data:-** 1.786943

**Varience of weight in given data:-** 3.193166

**Range of weight in given data:-** 8.399999999999999

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?

Expected value of weight of patient= weight of all patients/total no. of patients.

(108+110+123+134+135+145+167+187+199)/9

=145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

1. **To find skewness for cars speed and distance data:-**

* upload that file in the jupyter notebook.
* After that import pandas as pd and read that csv file.
* After displaying csv file –
* df.skew() is used to find the skewness-

Index 0.000000

speed -0.117510

dist 0.806895

dtype: float64

1. **To find kurtosis for cars speed and distance data:-**

* df.kurt() is used to find kurtosis –

Index -1.200000

speed -0.508994

dist 0.405053

dtype: float64

**SP and Weight(WT)**

**Use Q9\_b.csv**

1. **To find skewness for cars speed and weight data:-**

* Upload that file in the jupyter notebook.
* After that import pandas as pd and read that csv file.
* After displaying csv file –
* df.skew() is used to find the skewness-

Unnamed: 0 0.000000

SP 1.611450

WT -0.614753

dtype: float64

1. **To find kurtosis for cars speed and weight data:-**

* Upload that file in the jupyter notebook.
* After that import pandas as pd and read that csv file.
* After displaying csv file –
* df.kurt() is used to find the kurtosis-

Unnamed: 0 -1.200000

SP 2.977329

WT 0.950291

dtype: float64

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



ANS: In given histogram there is chick weight on the X-axis and frequency for the data is on the Y-axis. The scale for X-axis is 100 and the scale for Y-axis is 50. There is very less data on the range 300 and 400. The nature of skewness in given histogram is positive skewness so the large number of the data is present on the right side of the median.



ANS: In given boxplot mean is greater than median so the nature of skewness is positive skewness. As there positive skewness the large number of data is present on the right side of the boxplot. There are outliers on the upper extreme. As mean is greater than median the skewness is unsymmetric.

**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?

* Given data= n=2000

Sample weight=200

Standard derivation=30

* Confidence interval for 94%

=1.89

* Confidence interval for 98%

=2.32

* Confidence interval for 96%

=2.06

* (200-1.89\*(30/math.sqrt(2000)))

=198.73

* (200+1.89\*(30/math.sqrt(2000)))

=201.26

* (200-2.32\*(30/math.sqrt(2000)))

=198.44

* (200+2.32\*(30/math.sqrt(2000)))

=201.55

* (200-2.06\*(30/math.sqrt(2000)))

=198.61

* (200+2.06\*(30/math.sqrt(2000)))

=201.38

**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 of scores obtained by student- 41
   * Median of scores obtained by student- 40.5
   * Varience of scores obtained by student- 25.529411764705884
   * Standard derivation of scores obtained by student- 5.05266382858645

​

1. What can we say about the student marks?

ANS: in given data of students there are total 18 students with their marks.To find mean firstly we have to use that the data sequentially and if the data is not in sequence then we have to arrange it. But here data is already in sequential manner so we can find mean of that data directly.The median of the given data is 40.5 means it present at centre.So we can say maximum number of data is present on the centre so the given data is normally distributed.

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

* If mean and median of the given data are equal to each other then it is symmetrical and the nature of skewness is zero.

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

* For the given data if the mean is greater than median then it means that large number of data is present on right side and the skewness is positive.

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

* For the given data if the mean is less than median then the large number of data should be present on negative side and the skewness is negative.

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

* For the data positive kurtosis value indicates that it has heavy tailed data. Also we can say the more data is on the centre.

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

* For the data negative kurtosis value indicates that it has lighter tail and the the data is not at centre the data should be distributed.

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



* What can we say about the distribution of the data?

ANS: Here Q1 of the data is 10, Q3 of the data is 18 and Q2 is 14.6. Then IQR of the data is 8. 25% of the data is on Q1, 50% data Is on Q2 and 75% data is on Q3.

* What is nature of skewness of the data?

ANS: Here the large number of data is present on the left hand side so the nature of skewness is negative.

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

ANS: To find IQR of the data given in boxplot we use,

IQR= (Q3-Q1)

=18-10=8

Q19) Comment on the below Boxplot visualizations?

ANS: There are total two boxplots given in the question.The first boxplot has less data as compare to the boxplot 2. Both two of boxplots are symmetric because mean and median of the data are equal.



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

ANS: In comparison of both boxes boxplot 1 has less data than the boxplot 2.The data in boxplot 1 scattered between (230 to 280) where the data in boxplot is more and it is scattered between (100 to 350).

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)

1-p(38,loc=34.422076,scale=9.131445)

= 0.34759394041453007

= 34.75%

* 1. P(MPG<40)

P(40,loc=34.422076,scale=9.131445)

= 0.7293498604157946

= 72.93%

* 1. P (20<MPG<50)

P(50,loc=34.422076,scale=9.131445)-p(20,loc=34.422076,scale=9.131445)

= 0.9559926858516099- 0.05712377822429007

= 0.8988689076273199

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

ANS: After plotting the graph we get the result as large number of the data is present on line so we can say it is normally distributed.

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

Dataset: wc-at.csv

ANS: For this csv file as a result large number of data is present at the centre. So we can called it as symmetric data or we can also say it is normally distributed.

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

* Z score for 90% confidence interval:-

(1+0.90)/2

= 0.95

(1.6+0.05)

=1.65

* Z score for 94% confidence interval:-

(1+0.94)/2

= 0.97

(1.8+0.09)

=1.65

* Z score for 60% confidence interval:-

(1+0.60)/2

= 0.8

(0.8+0.05)

=0.85

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

* T score for 95% confidence interval with sample size is 25:-

First off all alpha= (1-0.95)

= 0.05

Now, to find t score we have to use:-

Stats.t.ppf(1-0.05/2,24)

=2.06389

* T score for 96% confidence interval with sample size is 25:-

First off all alpha= (1-0.96)

= 0.04

Now, to find t score we have to use:-

Stats.t.ppf(1-0.04/2,24)

=2.1715

* T score for 99% confidence interval with sample size is 25:-

First off all alpha= (1-0.99)

= 0.01

Now, to find t score we have to use:-

Stats.t.ppf(1-0.01/2,24)

=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

**ANS:-**

**Given:-**

**average mean=270**

**N=18**

**Standard derivation=90**

**Sample mean =260**

**So, we have to use formula:-**

**(sample mean-average mean)/(stdderivation/math.sqrt(18))**

**So by putting values in this formula we will get:**

**(260-270)/(90/math.sqrt(18))**

**= -0.4714**

**To find critical value:-**

**Stats.t.ppf(1-0.05,17)**

**= 1.7396067260750672**