|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete Data |
| Results of rolling a dice | Discrete Data |
| Weight of a person | Continuous Data |
| Weight of Gold | Continuous Data |
| Distance between two places | Continuous Data |
| Length of a leaf | Continuous Data |
| Dog's weight | Continuous Data |
| Blue Color | Nominal Data |
| Number of kids | Discrete Data |
| Number of tickets in Indian railways | Discrete Data |
| Number of times married | Discrete Data |
| Gender (Male or Female) | Nominal Data |

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

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

Answer--

If three coins are tossed, the probability of getting the two heads and one tail are:

Total output= (H,H,H), (H,H,T), (H,T,T), (H,T,H), (T,H,H), (T,T,H), (T,H,T), (T,T,T)

Total number of outcomes = 8

Total number of outcomes having two heads and one tail = 3

Probability of getting two heads and one tail = 3/8

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

1. Equal to 1

Answer—

When two dice are rolled the outcomes are:

(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), (5,6), (5,5), (6,1), (6,2), (6,3), (6,4), (6,5), (6,6)

Total number of outcomes = 36

Total number of outcomes for sum equals to one = 0

So the probability of getting sum equals to one = 0/36

1. Less than or equal to 4

Answer—

Total number of outcomes = 36

Total number of outcome for less than or equal to 4 = 6

(1,1), (1,2), (1,3), (2,1), (2,2), (3,1)

So the probability of getting less than or equals to 4 = 6/36 = 1/6

1. Sum is divisible by 2and 3

Answer :- The possible outcomes whose sum is divisible by both two and three are:

(1,5)

(2,4)

(3,3)

(4,2)

(5,1)

(6,6)

Therefore, the number of favourable outcomes is 6. The total number of possible outcomes is 36. Hence, the probability of getting a sum divisible by both two and three is:

P(sum divisible by both two and three) = Number of favourable outcomes / Total number of possible outcomes

                                      =6/36

                                      =0.166.

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?

Answer—

Formula used:- nCr=n!/(n-r)!\*r!

Probability for two balls selected is 7C2= 7!/5!\*2! = 7\*6\*5\*4\*3\*2\*1/ 5\*4\*3\*2\*1\*2 =21

Outcome for excluding two blue balls and other two balls selected= 2 red balls + 3 green balls = 5 balls

Therefore 5C2= 10

Probability P(X) = favorable outcomes/total outcomes = 10/21

Hence, probability two balls selected except blue ball = 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

Answer—

Number of child A can expect = 1\*0.015= 0.015-------eq-(1)

Number of child B can expect= 4\*0.20= 0.8-------eq-(2)

Number of child C can expect= 3\*0.65= 1.95-------eq-(3)

Number of child D can expect= 5\*0.005= 0.025-------eq-(4)

Number of child E can expect= 6\*0.01= 0.06-------eq-(5)

Number of child F can expect= 2\*0.120=0.24-------eq-(6)

Adding above equation (1-6 )= 0.015+0.08+1.95+0.025+0.06+0.24= 3.090

Adding all the above equation will give the expected number of candies for a randomly selected child that is 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**

Answer-

Mean-

Point- data[‘point’].mean()= 3.596563

Score- data[‘score’].mean()= 3.17250

Weigh- data[‘weigh’].mean()= 17.848750

Median

Point- data[‘point’].median()= 3.695

Score- data[‘score’]. median ()= 3.325

Weigh- data[‘weigh’]. median ()= 17.710

Mode

Point- data[‘point’].mode()= 3.07, 3.92

Score- data[‘score’]. mode ()= 3.44

Weigh- data[‘weigh’]. mode ()= 17.02

Standard Deviation

Point- data[‘point’].std()= 0.534679

Score- data[‘score’].std()= 0.978457

Weigh- data[‘weigh’].std()=1.786943

Variance

Point- data[‘point’].var()= 0.285881

Score- data[‘score’]. var()= 0.957379

Weigh- data[‘weigh’]. var()=3.193166

Inference- As Mean is closer to the Median and Mode, we can say that the data is relatively normally distributed.

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?

Answer-

The expected value of the weight of a random patient will be the average of all the 9 patients =(108+110+123+134+145+167+187+199)/9 = 1308/9 =145.33

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

**Cars sspeed and distance**

**Use Q9\_a.csv**

Answer—

**Speed--** We got kurtosis an skewness 0.51 and 0.12 respectively (from ipynb file). We can infer that distribution is slightly left skewed with longer tail on left and for kurtosis it is Platykurtic. There are fewer extreme value or outliers in the data set with lower tail ends compared to normal distribution.

**Distance—** We got skewness= 0.81 and kurtosis= 0.41 from that we can infer that skewness value is slightly positive, we can say distribution is slightly right skewed and longer tail on right end and for kurtosis it is Leptokurtic that is there are many extreme value or outliers in the dataset with heavy tails compared to the normal distribution.

**SP and Weight (WT)**

**Use Q9\_b.csv**

**Answer—**

**SP-- We** got skewness= 1.61 and kurtosis= 2.98 from that we can infer that skewness value is slightly positive, we can say distribution is slightly right skewed and longer tail on right end and for kurtosis it is Leptokurtic that is there are many extreme value or outliers in the dataset with heavy tails compared to the normal distribution.

**WT--** We got skewness= -0.61 and kurtosis= 0.95 from that we can infer that skewness value is slightly negative, we can say distribution is slightly left skewed and longer tail on right end and for kurtosis it is Leptokurtic that is there are many extreme value or outliers in the dataset with heavy tails compared to the normal distribution.

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



1. Histogram plot on Chick weight- From the above histogram we can observe that the distribution is right skewed distribution with more chicks weighing in the range of 50 to 150. So we can say Mean>Median>mode.



1. Boxplot- From the above boxplot we can observe that it is right skewed distribution and median is closer to the first quartile.

**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?

Answer— Sample of mean x = 200

Sample of standard deviation ∝ = 30

Sample size of n = 2000

Degree of freedom = 2000 – 1 = 1999,

t = 1.8916

Confidence interval for 94% is (198.73,201.27)

200-1.8916/30√2000 = 198.73

200+1.8916/30√2000 = 201.27

Confidence interval for 96% is

Degree of freedom = 199, t= 2.0673

200-2.0673/30√2000 = 198.61

200+2.0673/30√2000 = 201.39

Confidence interval for 98% is

Degree of freedom = 199, t= 2.3452

200-2.3452/30√2000 = 198.43

200+2.3452/30√2000 = 201.57

**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.

Answer—

**Mean**= Sum of marks of all 18 students /18

= 34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56/18

= 697/18 = 38.72

**Median** = First arrange the above data in Ascending order.

34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56

The middle value is 8TH marks and 9th marks so 40 and 41

Median = 40+41=81

81/2= 40.5

**Variance =** Formula used— We can calculate using the mean

((34-38.72)+(36-38.72)+(36-38.72)+(38-38.72)+(38-38.72)+(39-38.72)+(39-38.72)+(40-38.72)+(40-38.72)+(41-38.72)+(41-38.72)+(41-38.72)+(41-38.72)+(42-38.72)+(42-38.72)+(45-38.72)+(49-38.72)+(56-38.72))^2/18

=30.33

**Standard deviation**

STD = Sq. root of variance = Sq. root of 30.33 = 5.5066

1. What can we say about the student marks?

Answer— We find that the mean is38.72 and median is 40.5 which is almost close. From this we can say that the marks of the students are symmetrically distributed. Similarly from the variance and standard deviation we can find the marks are moderately variable.

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

Answer— When Mean=Median=Mode, we can say that data set is approximately symmetrically distributed. It implies that skewness is zero skewed.

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

Answer—When the mean > median, it indicates that the data is right skewed or positively skewed.

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

Answer— When the Median > Mean, it means that the data set is negatively distributed or we can say that the data is left skewed.

Q16) What does positive kurtosis value indicates for adata ?

Answer—Positive Kurtosis value indicates that the data is normally distributed and kurtosis is 0.

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

Answer—Negative kurtosis is also known as Platykurtic in which data has shorter tail and has fewer extreme values along with outliers. Flat peaks are often seen in this type of distribution. The dispersed data will be also less concentrated along mean.

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



What can we say about the distribution of the data?

Answer—In the above boxplot data we can say that data is heavily distributed along the left side of the plot.

What is nature of skewness of the data?

Answer—In the above boxplot, it is left skewed since the median is located much near to third quartile.

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

Answer—In the above boxplot we can see that Q1 is approx 10 and Q3 is nearly 18, so the IQR will be the difference of Q3 and Q1 that is 18-10 = 8. So IQR will be 8

Q19) Comment on the below Boxplot visualizations?



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

Answer—

From the boxplot 1 we get that:

Minimum- 237

Quartile 1(Q1) = 250

Quartile 3(Q3) = 275

Maximum = 287

From above median will be = (250+275)/2 = 262.5

From the boxplot 2 we can get that:

Minimum- 200

Quartile 1(Q1) = 225

Quartile 3(Q3) = 300

Maximum = 337

From above median will be = (225+300)/2 = 262.5

Hence from the above two boxplot distribution we can observe the following:

1. The data distribution is more in boxplot 2 and less in boxplot 1.
2. The median for both the boxplot is same that is 262.5 but IQR range of boxplot1 is 25 and for boxplot2 is 75 which is 3 times of boxplot1.
3. Also we can say that data is not skewed.

Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

Calculate the probability of MPG ofCars for the below cases.

MPG<- Cars$MPG

* 1. P(MPG>38)

Answer— From the IPYNB file we got the mean of MPG= 34.422and std deviation as 9.1314

So , P(MPG>38)= 1-0.65= 0.35=35%

* 1. P(MPG<40)

Answer— From the IPYNB file we got the mean of MPG= 34.422and std deviation as 9.1314

So , P(MPG<40) is 72.9%

* 1. P (20<MPG<50)

Answer— From the IPYNB file we got the mean of MPG= 34.422and std deviation as 9.1314

So , P(20<MPG<50) is 89.88%

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Answer— Refer to IPYNB file

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

Dataset: wc-at.csv

Answer— Refer to IPYNB file

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

Answer— Refer to IPYNB file

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

Answer— Refer to IPYNB file

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

Answer— Refer to IPYNB file