|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete |
| Results of rolling a dice | Discrete |
| Weight of a person | Continues |
| Weight of Gold | Continues |
| Distance between two places | Continues |
| Length of a leaf | Continues |
| Dog's weight | Continues |
| 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:

Three coins are tossed.

Sample space: {HHH,HHT,HTH,THH,TTH,THT,HTT,TTT}

Total no. of outcomes = 8

outcomes of getting two head and one tail are = {HHT,HTH,THH}

No. of outcomes = 3

Probability of getting two head and one tail = 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:

Two dices are rolled.

No. of Sample space = 6\*\*2 = 36

a) No. of outsomes of getting the sum as 1 = 0

Probability = 0

b) Outcomes of getting the sum less than equal to 4 are ={(1,1),(1,2),(1,3),(2,1),(2,2),(3,1)}

No. of outcomes = 6

Probability = 6/36 = 1/6

c)Outcomes of getting sum is divisible by 2 & 3 are = {(1,5),(2,4),(3,3),(4,2),(5,1),(6,6)}

No of outcomes = 6

Probability = 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 that none of the balls drawn is blue = 5C2/7C2 = 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

Ans:

Expected number of candies for a randomly selected child =   
0.015\*1+0.20\*4+0.65\*3+0.005\*5+0.01\*2+0.120\*2 = 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.

Ans:

|  |  |  |  |
| --- | --- | --- | --- |
|  | Points | Score | Weigh |
| Count | 32 | 32 | 32 |
| Mean | 3.5965 | 3.2172 | 17.8487 |
| Median | 3.695 | 3.325 | 17.710 |
| Mode | 3.07,3.92 | 3.44 | 17.02,18.90 |
| Variance | 0.2858 | 0.9573 | 3.1931 |
| Std | 0.5346 | 0.9784 | 1.7869 |
| Range | 2.760 – 4.930 | 1.513 – 5.424 | 14.50 – 22.90 |

tUse 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?

Ans:

Expected Value of the Weight of that Patient = (108, 110, 123, 134, 135, 145, 167, 187, 199)/9 = 145.333

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

Cars speed and distance

Use Q9\_a.csv

Ans:

|  |  |  |
| --- | --- | --- |
|  | Car\_speed | Distance |
| Skewness | -0.1175 | 0.80 |
| kurtosis | -0.5089 | 0.4050 |

The Skewness of Car\_speed is -ve ,it indicates the data is left skewed and mean<median.

The Skewness of Distance is +ve, which indicates the data is right skewed and mean>median.

The kurtosis for Car\_speed is -ve which indicates less kurtosis that means data have flat peak and thin tails.

The kurtosis for Distance is +ve which indicates excess kurtosis that means data have sharper peak and thick tails.

SP and Weight(WT)

Use Q9\_b.csv

Ans :

|  |  |  |
| --- | --- | --- |
|  | SP | Weight |
| Skewness | 1.6114 | -0.614 |
| kurtosis | 2.977 | 0.950 |

The Skewness of Weight is -ve ,it indicates the data is left skewed and mean<median.

The Skewness of SP is +ve, which indicates the data is right skewed and mean>median.

The kurtosis for SP is +ve which indicates excess kurtosis that means data have sharper peak and thick tails. And also we are seeing that kurtosis is 2.97 which is nearly equal to 3 that means data is distributed equivalent to its Normal distribution.

The kurtosis for Distance is +ve which indicates excess kurtosis that means data have sharper peak and thick tails.

Q10) Draw inferences about the following boxplot & histogram



Ans:

From the histogram we can observe that most of the observations concentrated on the left side, that means our data is right skewed.



Ans:

From this boxplot we can conclude that data concentrated on the left side of the box value that’s why it is right skewed and there are some outliers after the max value.

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: Confidence Interval for 94% = 198.73,201.26

Confidence Interval for 98% = 198.43,201.56

Confidence Interval for 96% = 198.62,201.37

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.
2. What can we say about the student marks?

Ans: 1)

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | mean | median | mode | variance | Std |
| Student\_score | 41.117 | 41.0 | 41 | 26.860 | 5.182 |

2)We can see that mean = median = mode here.

So we can conclude that data is normally distributed and perfectly symmetrical.

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

Ans: when mean = median , such kind of data is normally distributed as well as perfectly symmetrical so skewness will be 0.

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

Ans: when mean>median , data is going to be right skewed and skewness will be +ve.

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

Ans: when median>mean , data is going to be left skewed & skewness will be -ve.

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

Ans: +ve kurtosis value indicates that our data distribution will have sharper peak and thick tails.

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

Ans: -ve kurtosis value indicates that our data distribution will have flat peak and thin tails.

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



What can we say about the distribution of the data?

Ans:

In this we can see min is approximately 1 and max is approximately 19.

50% of data lies between 10 to 18.

Range is 19 – 1 = 18

First Quartile (Q1)= 10

Third Quartile (Q3)= 18

IQR = 18 – 10 = 8

Median is approximately 15.5

75% data lies between 10 to 19.

What is nature of skewness of the data?

Ans:

We can see that spreads for the box is more on the left side while concentration is more on the right side so data is left skewed.

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

First Quartile (Q1)= 10 ,Third Quartile (Q3)= 18

IQR = 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.

Ans:

The min value for boxplot 1 is approximately 243, max is approximately 290.

The min value for boxplot 2 is approximately 190, max is approximately 350.

Range for boxplot 1 = 290 – 243 = 47

Range for boxplot 2 = 350 – 190 = 160

That means spread for boxplot 2 is higher than boxplot 1 , that means boxplot 2 has more variability.

Median for both the boxplot is approximately 262.

Both the boxplots shows that data is normally distributed .

For boxplot 1 IQR = 284 – 260 = 24

For boxplot 2 IQR = 309 – 220 = 89

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)

Ans: 1-stats.norm.cdf(38,loc=mean ,scale=std)

0.3475

* 1. P(MPG<40)

Ans: stats.norm.cdf(40,loc=mean ,scale=std)

0.7293

* 1. P (20<MPG<50)

Ans: stats.norm.cdf(50, loc = mean ,scale = std) - stats.norm.cdf(20, loc =

mean ,scale = std)

0.898

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

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

Dataset: wc-at.csv

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

Ans: z score for 90% confidence interval = 1.64

z score for 94% confidence interval = 1.88

z score for 60% confidence interval = 0.841

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

Ans: stats.t.ppf(.975,24)

2.063

stats.t.ppf(.98,24)

2.171

stats.t.ppf(.995,24)

2.796

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