|  |  |
| --- | --- |
| 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. Favorable outcomes= HHT, THH, HTH= 3

Total outcomes= 23 = 8

Probability= 3/8 = 0.375

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. a) 0

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

6/36 = 0.166

1. sum divisible by 2 & 3 = 6, 12

P(6)=5/36, P(12)=1/36

5/36+1/36= 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?

Ans. Total no. of balls in bag= 7

Total outcomes= 7C2 = 7!/2!5! = 21

Favorable outcomes= 5C2

= 5!/2!3! = 10

P(no blue balls) = 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. Candies count=x

Probability= P(x)

Expected value= ∑ x P(x)

= (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**

Ans. Points:

Mean- 3.596563, Median- 3.695, Mode-3.07, 3.92, Range- 2.17, Variance- 0.285881, Standard Deviation- 0.534679

Score:

Mean- 3.21725, Median- 3.325, Mode- 3.44, Range- 3.911, Variance- 0.957379, Standard Deviation- 0.978457

Weigh:

Mean- 17.84875, Median- 17.71, Mode- 17.02,18.90, Range- 8.399, Variance- 3.193166, Standard Deviation- 1.786943

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.

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| X | 108 | 110 | 123 | 134 | 135 | 145 | 167 | 187 | 199 |
| P(X) | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 | 1/9 |

Expected Value= ∑X P(X)

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

= 145.33

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

**Cars speed and distance**

**Use Q9\_a.csv**

Ans. Speed- Skewness: -0.11751, Kurtosis: -0.50899

Inference: since the skewness of speed is negative the distribution is skewed towards left. Negative kurtosis means the peak of the distribution graph is flatter and the distribution is heavy tailed.

Distance- Skewness: 0.806895, Kurtosis: 0.405053

Inference: since the skewness of speed is positive the distribution is skewed towards right. Positive kurtosis means the peak of the distribution graph is steeper and the distribution is light tailed.

**SP and Weight(WT)**

**Use Q9\_b.csv**

Ans. SP- Skewness: 1.61145, Kurtosis: 2.977329

Inference: since the skewness of speed is positive the distribution is skewed towards right. Positive kurtosis means the peak of the distribution graph is steeper and the distribution is light tailed.

WT- Skewness: -0.61475, Kurtosis: 0.950291

Inference: since the skewness of speed is negative the distribution is skewed towards left. Positive kurtosis means the peak of the distribution graph is steeper and the distribution is light tailed.

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



Ans. Histogram- The distribution in the histogram is positively skewed.

Most of the data points lie in the range 50-100 with the frequency of

200.

Boxplot- There are many outliers present on the upper side of the box.

Lower Whisker has short length therefore less data points lie between Q0

And Q1. The boxplot shows that the distribution is positively skewed.

**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: N=3,000,000

n=2,000, x=200, s.d.=30

1)94% i.e., c=0.94

Significance level= 1-c= 1-0.94= 0.06

Df= n-1= 2000-1= 1999

t value= +- 1.881861

2)98% i.e., c=0.98

Significance level= 1-c= 1-0.98= 0.02

Df= n-1= 2000-1= 1999

t value= +- 2.328215

3)96% i.e., c=0.96

Significance level= 1-c= 1-0.96= 0.04

Df= n-1= 2000-1= 1999

t value= +- 2.05509

For 94% confidence interval is (198.118139, 201.881861)

For 98% confidence interval is (197.671785, 202.328215)

For 96% confidence interval is (197.94491, 202.05509)

**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.

|  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Scores | 34 | 36 | 38 | 39 | 40 | 41 | 42 | 45 | 49 | 56 |
| Frequency | 1 | 2 | 2 | 2 | 2 | 4 | 2 | 1 | 1 | 1 |

1. Mean= [(34\*1)+(36\*2)+(38\*2)+(39\*2)+(40\*2)+(45\*1)+(49\*1)+(56\*1)]/18

= 41

Median= (40+41)/2= 40.5

Variance= ∑ (x-mean)\*2/N-1 = 25.52

Standard Deviation= √variance = 5.05

1. The mean of the student marks is 41 with a deviation of -5.05 to +5.05

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

Ans. Symmetric

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

Ans. Positively Skewed (right)

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

Ans. Negatively Skewed (left)

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

Ans. In positive kurtosis the curve is steeper and the distribution is light tailed. More percent of data is present near the tail.

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

Ans. In negative kurtosis the curve is flatter and the distribution is heavy tailed. Few data are present near the tail.

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



What can we say about the distribution of the data?

Ans. The major amount of data lies between 10-18. The median i.e., Q2 lies between 14-16.

Lower Quartile, Q1= 10: 25% of data lies below this

Upper Quartile, Q3= 18: 75% of data lies below this

There are no outliers

What is nature of skewness of the data?

Ans. The distribution is skewed to the left i.e., median>mean.

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

Ans. From the box plot, Q1= 10 and Q3= 18

IQR= Q3-Q1 = 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 median of Boxplot 1 and 2 is same. Since in both the distributions 1 and 2, the data is in equal proportions about the median, therefore both are normally distributed i.e., symmetric.

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)
  2. P(MPG<40)
  3. P (20<MPG<50)

Ans. 1) P(MPG>38) = 0.3475

2) P(MPG<40) = 0.7293

3) P (20<MPG<50) = 0.8988

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans. MPG mean= 34.422

MPG median= 35.152

Since mean is not equal to median, MPG of cars does not follow normal distribution.

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

Dataset: wc-at.csv

Ans. AT mean= 101.894

AT median= 96.54

Since mean is not equal to median, AT does not follow normal distribution.

Waist mean= 91.901

Waist median= 90.8

Since mean is not equal to median, Waist does not follow normal distribution.

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

Ans. 90% confidence interval= 1.6448

94% confidence interval= 1.8807

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

Ans. 95% confidence interval= 2.0638

96% confidence interval= 2.1715

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

Ans. n=18, x bar=260, s=90, µ=270

df= n-1= 17

t= x bar- µ/ (s/√n) = (260-270)/ (90/√18) = -0.471

P(t≤-0.471) = 0.321 (using t distribution calculator and python)

There is a 32.1% chance that 18 randomly selected bulbs would have an average life of no more than 260 days.