Q1) Identify the Data type for the Following:

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

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 | Interval |
| Time on a Clock with Hands | Interval |
| Number of Children | Ratio |
| Religious Preference | Nominal |
| Barometer Pressure | Ratio |
| SAT Scores | Ordinal |
| Years of Education | Interval |

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

**Ans 3**: Since we know that one tails is obtained

The probablity are ={H,T,H},{T,H,H},{H,H,T}

Total probablity are ={H,H,H},{T,T,T},{H,T,H},{T,H,H},{H,H,T},{T,T,H}

{T,H,T},{H,T,T}

Probablitity= Favourable outcome = 3/8 (Answer).

Total outcomes

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

* Equal to 1
* Less than or equal to 4
* Sum is divisible by 2 and 3

**Ans 4:** a.) Probablitiy of sum equals to 1 is 0 .

b.) Probability of getting less than /equals to 4 = (2,1)(1,2)(3,1)(1,3)(2,2)(1,1) = 6

Total Probability = 6 x 6 =36

Probablitity= Favourable outcome = 6 = 1 (Answer).

Total outcomes 36 6

c.) Sum of dices is divisible by 2 and 3 i.e. 6 = (5,1)(1,5)(3,3)(2,4)(4,2)(6,6)

Total Probability = 6 x 6 = 36

Probablitity= Favourable outcome = 6 = 1 (Answer).

Total outcomes 36 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 5:** n = 7,r = 2

Probability to draw 2 balls out of 7 is 21.

Since we dont want to draw blue balls.

Probability to draw 2 balls out of 5 is 10

P(E) = 10/21 = 0.476

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 6.** Expected number of candies = 1 x 0.015 + 4 x 0.20 + 3 x 0.65 +

5 x 0.005 + 6 x 0.01 + 2 x 0.120

= 0.015+0.8+1.95+0.025+0.06+0.24

=3.09 (Answer).

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 7 :**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | MEAN | MEDIAN | MODE | Variance | Std. Dev | Range |
| Points | 3.59 | 3.70 | 3.92 | 0.29 | 0.53 | 2.17 |
| Score | 3.22 | 3.33 | 3.44 | 0.96 | 0.98 | 3.91 |
| Weigh | 17.85 | 17.71 | 17.02 | 3.19 | 1.79 | 8.40 |

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 8**: Probability of 1 patient selected = 1/9

Expected Value = (108 \* 1/9) + (110 \* 1/9) + (123 \* 1/9) + (134 \* 1/9) + (135 \* 1/9) + (145 \* 1/9) + (167 \* 1/9) + (187 \* 1/9) + (199 \* 1/9) = 145.33

Ans: 145.33 pounds(Answer)

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

**Cars speed and distance**

**Use Q9\_a.csv**

**SP and Weight(WT)**

**Use Q9\_b.csv**

**Ans 9.**

|  |  |  |
| --- | --- | --- |
|  | Skewness | Kurtosis |
| Speed | -0.11 | -0.57 |
| Distance | 0.78 | 0.24 |
| SP | 1.58 | -0.60 |
| Weight | 2.72 | 0.82 |

**Summary :** For Speed & Distance: Speed is slightly negatively skewed.While distance is positively skewed.Both have laptokurtic Distribution(+ve). There are outliers in distance.

For Speed is data is very concentrated in middle. FOr Distance data is concentrated on lower side of vaules.

For SP & Weight : Both SP & WT coulmn have outliers. WT is negatively skewed, while SP is positively skewed. Laptokurtic Distribution.

SP has lots of outlers and data is spread . While WT is very concentrated and has outliers on both sides.

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





**Ans 10**. **Fig(a):** We can say that data is positively skewed(leaning to left tail on right) and kurtosis is also positive. as there are few outliers( the curve is tall and narrow). Most weights are between 50-150(200 frequency) .weight 300- 400 is outliers.

**Fig(b):** From we can deduce that data is positively skewed as Mean is closer to bottom and whiskers are short on this side. It means data has few outliers. The size of box is also not large means data is not very dispersed

**Summary** : Assuming both figure represent same data. We can conclude that data is positively skewed and kurtosis is Laptokurtic(ie positive). The dispersion is less in data and there are few outliers.

**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 11**.

|  |  |  |
| --- | --- | --- |
| 94% | 96% | 98% |
| 198.73 | 198.44 | 198.62 |
| 201.26 | 201.56 | 201.37 |

The confidence interval for 94% is (198.738325292158, 201.261674707842)

The confidence interval for 96% is (198.43943840429978 ,201.56056159570022)

The confidence interval for 98% is (198.62230334813333 ,201.37769665186667)

**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 12.**

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | MEAN | MEDIAN | MODE | Variance | Std. Dev |  |
| Data | 41 | 40.5 | 41 | 25.53 | 5.05 |  |
| Score | 3.22 | 3.33 | 3.44 | 0.96 | 0.98 |  |

**Summary**: The marks are consistent across data. 56 is a outlier.

variance and deviation is high suggesting data is dispersion along the curve.

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

**Ans 13**. When mean =median means its a bell courve and data is normally distributed. hence, skewness is zero.

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

**Ans14**. Data is positively skewed.

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

**Ans15**. Data is negatively skewned.

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

**Ans16**. Positive kurtosis means thick tails and more frequent outliers.

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

**Ans 17**. Negative kurtosis means data is distributed.hence thin tail and less outliers compared to positive.

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



What can we say about the distribution of the data?

**Ans 18:** lower quartile =10,upper quartile =18, Mean=10, Median=15.

Data is more inclined towards right hence not symmetrical.There is no outliers as whole of data is covered by either box or whiskers. 50% of data lies above 10 and below 15. 75% below 18.

What is nature of skewness of the data?

**Ans:** Since mean > median data is negatively skewed. Also median is

towards the uppers limit and lower whisker is larger. Hence negatively skewed.

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

**Ans:** 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.

**Ans19:** Data is normally distributed**.** Box2 is more spread than Box1. Both data are perfectly skewed. Both have visibly same median. Upper and lower quartiles of both are different (box1 is smaller than box2) . There are potential outliers in box1, box 2 covers the whole of value range.

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 20**. P(MPG>38) = 0.347

P(MPG<40) = 0.729 = 0.73

P(20<MPG<50) = 0.898

Q 21) Check whether the data follows normal distribution

* Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

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

Dataset: wc-at.csv

**Ans 21**. a) MPG follows Normal Distribution.

b) Both AT and Waist doesnt follow Normal Distribution.

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

**Ans 22.**

|  |  |  |
| --- | --- | --- |
| 90% | 94% | 60% |
| 1.64 | 1.88 | 0.84 |

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

**Ans 23**.

|  |  |  |
| --- | --- | --- |
| 95% | 96% | 99% |
| 2.39 | 2.17 | 3.09 |

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 24**. Given,

s=260

S=270

samplestd=90

sample=18

t\_value =-0.471

The probability that on an avg bulb lasts less than 260 days is 0.321