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 | 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 |

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

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

Ans:- Sample Space= { HHH, HHT, HTH, HTT, THH, THT, TTH, TTT}

Probability=Interested Outcome/Total No. of Outcomes

= 3/8

= 0.375

Hence, When three coins are tossed, the probability of getting two heads and one tail is 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:- Sample Space = { 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}

Total No. of Events= 36

1. No. of Interested Events = 0

Probability of getting Sum equal to 1=0/36=0

1. No. of Interested Events = 6

(1 1, 1 2, 1 3, 2 1, 2 2, 3 1)--- sum Less than or equal to 4

Probability of getting sum Less than or equal to 4 = 6/36 = 1/6 =0.167

1. No. of Interested Events=6

( 1 5, 2 4, 3 3, 4 2, 5 1, 6 6)---divisible by 2 and 3 both

Probability of getting sum divisible by 2 and 3 = 6/36 = 1/6 =0.167

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:- Sample Space={Red Red, Red Green -6, Red Blue -4, Green Blue -6, Green

Green -3, Blue Blue}

Probability that none of the balls drawn is blue = 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 = 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:-

> mean(Points)

[1] 3.596563

> mean(Score)

[1] 3.21725

> mean(Weigh)

[1] 17.84875

> median(Points)

[1] 3.695

> median(Score)

[1] 3.325

> median(Weigh)

[1] 17.71

> Mode(Points,multi=T,warn = T)

[1] 3.07 3.92

> Mode(Score,multi=T,warn = T)

[1] 3.44

> Mode(Weigh,multi=T,warn = T)

[1] 17.02 18.90

> var(Points)

[1] 0.2858814

> var(Score)

[1] 0.957379

> var(Weigh)

[1] 3.193166

> sd(Points)

[1] 0.5346787

> sd(Score)

[1] 0.9784574

> sd(Weigh)

[1] 1.786943

> range(Ques7$Points)

[1] 2.76 4.93

> range(Ques7$Score)

[1] 1.513 5.424

> range(Ques7$Weigh)

[1] 14.5 22.9

Inferences:-

1. Mean, Median and Modal Values for the Data set provided slightly differ for Points, Score and Weigh Columns. Also we got bimodal values for Points and Weigh. No outliers are present in the data set. Hence, Mean is the best measure for Central Tendency for this Data Set.

2. There is less Variance for Points and Score but a lot of Variance for Weigh. Hence, Data is much disperse or widely spread for Weigh.

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 for Discrete Data is Mean = (108+110+123+134+135+145+167+ 187+199)/9

**Expected Value= 145.33pounds**

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

Cars speed and distance

Use Q9\_a.csv

Ans:-

> skewness(speed)

[1] -0.1139548

> skewness(dist)

[1] 0.7824835

> kurtosis(speed)

[1] 2.422853

> kurtosis(dist)

[1] 3.248019

SP and Weight(WT)

Use Q9\_b.csv

Ans:-

> skewness(Ques9b$SP)

[1] 1.581454

> skewness(Ques9b$WT)

[1] -0.6033099

> kurtosis(Ques9b$SP)

[1] 5.723521

> kurtosis(Ques9b$WT)

[1] 3.819466

**The negative skewness of variable implies that mass of distribution is concentrated on right whereas for positive skewness it is on left.**

**Kurtosis curve has high peak as values are positive.**

Q10) Draw inferences about the following boxplot & histogram



Ans:- Highest peak of Chicks lie between 50-100 weight. There is a constant decrease in frequency with increase in weight after 100.



Ans:- Median of the data is more towards the lower IQR. There are some outliers in the data. Most of the data is concentrated towards the lower limit.

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:- Sample size =2000

Degrees of Freedom (Df) = 2000-1= 1999

S/√n = 30/√(2000) = 0.6708

1. 94% Confidence Interval

**α = (1-0.94)=0.06**

**Critical Probability (p\*)=1-** **α/2= 0.97**

**T- score = 1.882**

**T-score \*** S/√n =1.26248

**Lower End of the Range = 200-**1.26248= 198.7375

**Upper End of the Range = 200+**1.26248= 201.26248

**Confidence Interval = [ 198.74 201.26 ]**

1. 98% Confidence Interval

**α = (1-0.98)=0.02**

**Critical Probability (p\*)=1-** **α/2= 0.99**

**T- score = 2.328**

**T-score \*** S/√n =1.56166

**Lower End of the Range = 200-**1.56166= 198.4383

**Upper End of the Range = 200+**1.56166= 201.56166

**Confidence Interval = [ 198.44 201.56 ]**

1. 96% Confidence Interval

**α = (1-0.96)=0.04**

**Critical Probability (p\*)=1-** **α/2= 0.98**

**T- score = 2.055**

**T-score \*** S/√n =1.37853

**Lower End of the Range = 200-**1.37853=198.62

**Upper End of the Range = 200+**1.37853= 201.37853

**Confidence Interval = [ 198.62 201.38 ]**

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.

Ans:- > mean(y)

[1] 41

> median(y)

[1] 40.5

> var(y)

[1] 25.52941

> sd(y)

[1] 5.052664

1. What can we say about the student marks?

Ans:- At an average, all the students scored around 41 marks in the tests. A small standard deviation indicates that all the data values are clustered closely around the mean.

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

Ans:- It would be **perfectly symmetrical distribution. No skewness.**

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

Ans:- The distribution is **positively skewed**.

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

Ans:- The distribution is **negatively skewed**.

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

Ans:- A distribution with a **positive kurtosis value indicates** that the distribution has heavier tails and a sharper peak than the normal distribution.

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

Ans:- A distribution with a **negative kurtosis value indicates** that the distribution has lighter tails and a flatter peak than the normal distribution.

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



What can we say about the distribution of the data?

Ans:- A distribution is considered "Negatively Skewed" when mean < median. It means the data constitute higher frequency of low valued scores

What is nature of skewness of the data?

Ans:- **Left Skewed or negatively skewed** because Median is more shifted towards third quartile(Q3) and the whisker is shorter on the upper end of the box.

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

Ans:- IQR= Q3-Q1=18 -10=**8(Approx)**   
  
  
Q19) Comment on the below Boxplot visualizations?



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

Ans:- 1] Median of Boxplot 1 = Median of Boxplot 2 = 262.5(Approx)

2] Boxplot 1 is **Right skewed or positively skewed** because median is slightly shifted towards Q1(first quartile), whereas Boxplot 2 shows **perfectly symmetrical distribution.**

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:- > mean(MPG)

[1] 34.42208

> sd(MPG)

[1] 9.131445

* 1. P(MPG>38)

> a=1-pnorm(38,34.4,9.13)

> print(a)

[1] 0.3466781

* 1. P(MPG<40)

> pnorm(40,34.4,9.13)

[1] 0.7301817

c. P (20<MPG<50)

> c=pnorm(50,34.4,9.13)-pnorm(20,34.4,9.13)

> print(c)

[1] 0.8988697

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

Ans:- > median(MPG)

[1] 35.15273

> mean(MPG)

[1] 34.42208

> kurtosis(MPG)

[1] 2.352262

Values for mean and median has slight difference. Also Kurtosis Value is slightly positive. Hence, MPG for Cars follows 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:- > mean(WT$Waist)

[1] 91.90183

> median(WT$Waist)

[1] 90.8

Value of Mean is not so far from Median. Hence, we can say that Waist Circumference(Waist) follows a normal distribution.

> mean(WT$AT)

[1] 101.894

> median(WT$AT)

[1] 96.54

There is a remarkable difference between Mean and Median. Hence, Adipose Tissue (AT) does not follow normal distribution.

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

Ans:- For 90% CI-

> qnorm(0.95)

[1] 1.644854

For 94% CI-

> qnorm(0.97)

[1] 1.880794

For 60% CI-

> qnorm(0.80)

[1] 0.8416212

Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99%

confidence interval for sample size of 25

Ans:- Df=25-1=24

T-score for 95% CI-

**α = (1-0.95)=0.05**

**Critical Probability (p\*)=1-** **α/2= 0.975**

**T- score = 2.064**

T-score for 96% CI-

**α = (1-0.96)=0.04**

**Critical Probability (p\*)=1-** **α/2= 0.98**

**T- score = 2.172**

T-score for 99% CI-

**α = (1-0.99)=0.01**

**Critical Probability (p\*)=1-** **α/2= 0.995**

**T- score = 2.797**

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:- Population mean = 270 days

n = 18

df(Degree of Freedom)= n-1= 17

Sample mean = 260

Sample deviation = 90 days

T-score = -0.471

**Probability = 0.321814**

> pt(-0.471,17)

[1] 0.321814