**Basic Statistics\_Level 1**

|  |  |
| --- | --- |
| 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 | Continous |
| 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 |

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 | Interval |
| Time on a Clock with Hands | Ratio |
| Number of Children | Ordinal |
| Religious Preference | Ordinal |
| 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** : X={HHH, HHT, HTH, HTT, THH, THT, TTH, TTT }

P(X= 2Heads and 1 Tail) = { HHT, HTH, THH}

= 3/8

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

1. Equal to 1

– **Ans**. 0

1. Less than or equal to 4

– **Ans**. 1/6

1. Sum is divisible by 2 and 3

**Ans**. 1/36

**Explanation** : X= {(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) }

1. Equal to 1 :

P(X=sum equal to 1) = 0/36 = 0

1. Less than or equal to 4 :

P(X= Less than or equal to 4) = {(1,1), (1,2), (1,3), (2,1), (2,2), (3,1}

= 6/36 = 1/6

1. Sum is divisible by 2 and 3 :

P(X= )= { (1,5), (2,4), (3,3), (4,2), (5,1), (6,6) }

= 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**.. P(=!Blue)= (5/7)\*(4/6)= 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 = 1\*0.015)+(4\*0.20)+(3\*0.65)+(5\*0.005)+(6\*0.01)+(2\*0.12) =3.09 or 3.

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** : Mean of points, Score, Weigh are 3.59 , 3.21 , 17.84 respectively

Median of points, Score, Weigh are 3.695, 3.325, 17.71 respectively

Mode of points, Score, Weigh are 3.92, 3.44, 17.02 respectively

Variance of points, Score, Weigh are 0.285, 0.957, 3.19 respectively

Standard deviation of points, Score, Weigh are 0.52, 0.96, 1.758801 respectively

Range of points, Score, Weigh are 2.17 , 3.911, 8.4 respectively

Observation : Mean & Median of Points, score, weigh are nearby which shows no presence of outliers

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** : Probability of any random patient = 1/9

Expected value (Weight of patient)=1/9\*(108+110+123+134+135+145+167+187+199)=436/3= 145.33

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

**Cars speed and distance**



**Ans** :

Speed :

kurtosis(speed) = 2.99 , skewness(speed)= -0.845

* As the kurtosis of ‘speed’ data is positive, it has thin peak and negligible tail. Also, negative skewness of ‘speed’ data implies that mass of the distribution of the data is concentrated on the right.

Dist :

kurtosis(dist) = 4.82, skewness(dist)= 1.22

* As the kurtosis of ‘dist’ data is positive, it has thin peak and negligible tail. Also, positive skewness of ‘dist data implies that mass of the distribution of the data is concentrated on the left.

**SP and Weight(WT)**



**Ans** :

SP :

kurtosis(SP) = 2.09 , skewness(SP)= -0.408

* As the kurtosis of ‘SP data is positive, it has thin peak and negligible tail. Also, negative skewness of ‘speed’ data implies that mass of the distribution of the data is concentrated on the right.

**Dist** :

kurtosis(WT) =3.82, skewness(WT)= -1.29

* As the kurtosis of ‘WT’ data is positive, it has thin peak and negligible tail. Also, neagtive skewness of ‘dist’ data implies that mass of the distribution of the data is concentrated on the right.

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



**Ans** : From the boxplot and histogram, we can say that the data is right skewed and it has 5 or more outliers. Majority of the data is concentrated around 0-200.

Also, we can say that Mean>Median means a neagtive skewness of data which implies that mass of the distribution of the data is concentrated on the right.

**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** : Average weight (94% confidence interval) = 1.26094

Average weight (98% confidence interval) = 1.55976

Average weight (96% confidence interval) = 1.37687

**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= 41, Mode= 41, Median= 40.5

1. What can we say about the student marks?

**Ans**: From the students marks, we can say that the average marks a student got is 41 and 50% of the students marks lie above and below 40.5 respectively. The number of students who got 41 marks is more than rest of the marks.

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

**Ans** : When mean, median and mode of the data is equal, it means that the data has skewness = 0 and the data is normally distributed.

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

**Ans** : Mean>Median means a neagtive skewness of data which implies that mass of the distribution of the data is concentrated on the right.

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

**Ans** : Median>Mean means a positive skewness of data which implies that mass of the distribution of the data is concentrated on the left.

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

**Ans** : The positive kurtosis of data means that it has thin peak and negligible tail.

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

**Ans** : The negative kurtosis of data means that it has wide peak and thinner tail.

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



What can we say about the distribution of the data?

**Ans** : From the box plot, we can say that the date isn’t normally distributed and the mean of the data is greater than the median. The middle % of the data falls within the range 10 to 18.

What is nature of skewness of the data?

**Ans**: The nature of skewneess of the data is neagtive which implies that mass of the distribution of the data is concentrated on the right.

What will be the IQR of the data (approximately)?   
  
 **Ans** : From the box plot : Q3=18 , Q1=10

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** : From the box plot 1, we can say that the data vaules fall within 240 to 280. 240 is the lower extreme and 280 is the upper extreme. The mean of the data is around 265. It is normally distributed.

From the box plot 2, we can say that the data vaules fall within 195 to 350. 195 is the lower extreme and 350 is the upper extreme. The mean of the data is around 265. It is normally distributed.

Both box plot 1 and box plot have the same mean and different upper and lower extremes. The spread of the data is more in box plot 2 than box plot 1. The data of both box plot 1 and box plot 2 are normally distributed.

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** : using pnorm(),

1. P(MPG>38) = 0.0728
2. P(MPG<40) = 0.0645
3. P (20<MPG<50) = 0.146

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Ans** : The skewness of the Cars dat set is -0.175 which mmeans that the data is not normally distributed. It is negatively skewed.

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

Dataset: wc-at.csv

**Ans**: Both Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set doesn’t follow Normal Distribution. The data is positively skewed.

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

**Ans** : Using qnorm(),

Z scores of 90% confidence interval = 1.64

Z scores of 94% confidence interval = 1.88

Z scores of 60% confidence interval = 1.04

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

**Ans** : T score of 95% confidence interval of 25 = 2.391

T score of 96% confidence interval of 25 = 2.492

T score of 99% confidence interval of 25 = 2.064

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** : 8.62e-18