|  |  |
| --- | --- |
| 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 | Nominal |
| 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 | Continuous |
| Hair Color | Nominal |
| Socioeconomic Status | Nominal |
| Fahrenheit Temperature | Ratio |
| Height | Continuous |
| Type of living accommodation | Nominal |
| Level of Agreement | Ordinal |
| IQ(Intelligence Scale) | Continuous |
| Sales Figures | Ratio |
| Blood Group | Nominal |
| Time Of Day | Ratio |
| Time on a Clock with Hands | Ratio |
| Number of Children | Discrete |
| Religious Preference | Nominal |
| Barometer Pressure | Continuous |
| SAT Scores | Continuous |
| Years of Education | Ordinal |

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

Total possibility = 2^3 = 8 [(HHH),(TTT),(TTH),(THT),(HTT),(HHT),(HTH),(THH)]

Getting 2 heads and 1 tail is 3/8

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

1. Equal to 1 – Whne 2 dyes are rolled the least u/p is (1,1) So getting o/p 1 is 0
2. Less than or equal to 4

– Total events = 6^2

Interested events is <= 4 [11, 12, 13, 21, 22, 31 ] = 6/36=1/6

1. Sum is divisible by 2 and 3

* Total events = 6^2 = 36
* Interested events is divisible by 2 & 3, All 6 factors are divisible by both 2 and 3 [15, 24, 33, 42, 51, 66] 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?

Here we go with combinations

NcR = n! / R! (N-R)!

Probability = Interested events / Total events

Total events = picking 2 balls from 7 balls = 7c2 = 7 \* 6 / 2! = 21

Interested events = Not picking blue balls, only 2 balls from 2 red and 3 green = 5c2 = 5\*4/2=10

Probability = 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

Expected value with probability distribution is ∑ X p(x)

Expected number of candies for a randomly selected child = (1\*0.015)+(4\*0.2)+(3\*0.65)+(5\*0.005)+(6\*0.01)+(2\*0.120)

0.015+0.8+1.95+0.025+0.06+0.24

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.

Mean-population µ = ∑ X / N sample – x = ∑ x / n

Median – sort the series and middle number is median if there are 2 middle numbers take the avg of 2 numbers gets median

Mode – Highest frequency number in a series, there can be more than 1 mode in a series

Variance – Spread of data is variance σ^2 = ∑ (X-µ)^2 / N

Standard Deviation – deviation from mean or simply square-root of variance

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Points** | **Score** | **Weight** |
| **sum** | 115.09 | 102.952 | 571.16 |
| **Mean** | 3.5965625 | 3.21725 | 17.84875 |
| **Median** | 3.695 | 3.325 | 17.71 |
| **Mode** | 3.07 & 3.92 | 3.44 | 17.02 & 18.9 |
| **Varience** | 0.276947559 | 0.927460875 | 3.093379687 |
| **Standard Deviation** | 0.526258072 | 0.963047701 | 1.758800639 |
| **Range** | 2.17 | 3.911 | 8.4 |

Inferences:

All the above 3 are not under symmetrical distribution



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?

Expected value without probability distribution is just its MEAN

145.33

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

**Cars speed and distance**



**SP and Weight(WT)**



|  |  |  |
| --- | --- | --- |
|  | Speed | Distance |
| Skewness | -0.89542 | 1.290763 |
| Kurtosis | 0.249561 | 2.464546 |

Skewness: measures the asymmetry in data

Speed(Negatively skewed) , Distance (greater positive skewness)

Kurtosis: kurtosis is the used to identify outliers in data. Three types of kurtosis are Mesokurtic, Leptokurtic, platokurtic. Kurtosis of a normally distributed(Mesokurtic) data is 3, Formula is Kurtosis – 3

Speed(0.24 - 3) -platokurtic, Distance(2.46 - 3) - platokurtic

|  |  |  |
| --- | --- | --- |
|  | SP | WT |
| Skewness | -0.42675 | -1.35187 |
| Kurtosis | -0.86374 | 1.160802 |

Skewness: SP (Normally distributed Data), WT (Higher negative Skewness)

Kurtosis : Speed(0.86 - 3) -platokurtic, Distance(1.16 - 3) - platokurtic

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



Histogram is used to represent continous data in ranges, With this we cant identify outliers



BoxPlot : Identify outliers, with Entire data it shows it in 5 summaries, 1 minimum value, 2Maximum value, 3- Median, 1st Quarntile is min to first Median’s Median, 2nd Quarntile is Max to First median’s Median

**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 ?

Sample size = 2000 , Total population = 30,00,000

X=200, SD = 30

Interval Estimate or confidence interval = X +- Z(SD/square root(n))

squareroot(2000) = 44.721

94% = (z-score for 97), 200 +- 1.89(30/44.721)= 200 +- 1.267 = 198.733 to 201.267

98% =(z-score for .99), 200 +- 2.33(30/44.721)= 200 +- 1.563 = 198.437 to 201.563

96% =(z-score for .98), 200 +- 2.06(30/44.721)=200 +- 1.381 = 198.619 to 201.381

**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?

|  |  |
| --- | --- |
| Mean | 41.00 |
| Median | 40.50 |
| Varience (Sigma square) | 24.11111 |
| Standard Deviation (Sigma) | 4.910307 |

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

Data is in Normal Distribution, when Mean=Median=Mode

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

Data is positively skewed when mean > median

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

Data is negatively skewed when mean < median

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

Kurtosis study about sharpness/peakness of curve,it studies about outliers present in data

Positive kurtosis means Leptokurtic, kurtosis for a normally distributed data is 3. So kurtosis is always compared to Mesokurtic(i.e., 3) (Kurtosis -3)

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

Negative kutosis means platokurtic (kurtosis -3)

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



What can we say about the distribution of the data?

Median(50% in data) lies between 15 & 16

Lower quarantine(Q1-25%) = 10

Upper quarantine(Q3-5%) = 18

IQR (Inter Quarantile range) = Q1 to Q3 (25% to 75%) = 8

Minimum = Q1 – 1.5 (IQR) = 10-12 = -2

Maximum = Q3 + 1.5 (IQR) = 18+12=30

What is nature of skewness of the data?

Negatively skewed data

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

IQR (Inter Quarantile range) = Q1 to Q3 (25% to 75%) = 8

Q19) Comment on the below Boxplot visualizations?



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

Both the data is normally distributed

Same Median

Mean and Median are same for the datasets

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)

c. P (20<MPG<50)

Total events = 81

P(MPG > 38) = 33/81 = 0.407

P(MPG < 40) = 61/81 = 0.753

P(20 < MPG < 50) = 69/81 = 0.851

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

skewness(cars$MPG)

-0.1746343 For normal distribution skewness should be zero

kurtosis(cars$MPG)

2.352262 For normal distribution kurtosis ahould be 3, this is platykurtic

Find whether skewness is zero and kurtosis is 3 for a data, then they are in 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

> skewness(normalized\_data\_wc\_at)

Waist AT

0.1322042 0.5767897

> kurtosis(normalized\_data\_wc\_at)

Waist AT

1.892724 2.672812

No, Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set does not follows Normal Distribution

Draw a QQ-plot for the both and if most of the data falls on QQ-line(45 deg line). Then both are normally distributed to each other

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

Z-score = (x – mean)/sigma

For 90% check for .950 =1.65

94%-> .97 = 1.89

60% -> .80 = 0.85

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

> qt(0.95,24)

[1] 1.710882

When you don’t have standard devaiation (Sigma) then we go for T-scores

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

Population Mean = 270 days

Sample size = 18 bulbs

Sample Mean = 260 days

Samples standard deviation = 90 days

T-score formula = (sample mean – population mean) / (sample SD/squareroot(sample size) )

(260-270)/(90/squareroot(18) ) = -10/21.2134 = -0.4714

> pt(-0.4714,17)

[1] 0.3216741