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

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

Ans: Probability = No of interested outcomes/ Total no of outcomes

Total no of outcomes = {HHH, HHT, HTH, HTT, THH, THT, TTH, TTT} = 8

No of interested outcomes i.e., chances of getting two heads and a tail is 3.

Therefore, the probability will be 3/8.

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: Probability = No of interested outcomes/ Total no of outcomes

Total no of outcomes = { (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) } =36

1. Equal to 1: 0/36 = 0
2. Less than or equal to 4: 6/36 => 1/6
3. Sum is divisible by 2 and 3: 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: Total no of balls is = 7

2 balls are drawn at random out of 7 balls= 7C2 = nCr

i.e, 7! / (5!\*2!) =21

Probability that none of the balls drawn is blue = 5C2

i.e., 5! / (2!\*3!) = 10

Therefore, Probability is 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: To calculate the expected value, the formulae is given by:

Let candies count be x, Probability be P(x):

Expected value = ∑(P(x) \* x)

Therefore, the expected value will be:

EV= ((1\*0.015)+(4\*0.20)+(3\*0.65)+(5\*0.005)+(6\*0.01)+(2\*0.120))

EV=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.



Ans:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Points: |  | Weight: |  | Score: |  |
| Mean | 3.5965625 | Mean | 17.84875 | Mean | 3.215219 |
| Median | 3.655 | Median | 17.51 | Median | 3.2025 |
| Mode | 3.07 & 3.92 | Mode | 17.02 & 18.9 | Mode | 3.44 |
| Variance | 0.285881351 | Variance | 3.193166129 | Variance | 0.956627 |
| StdDev | 0.534678736 | StdDev | 1.786943236 | StdDev | 0.978073 |
| Range | 2.17 | Range | 8.4 | Range | 3.911 |

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: As the formula of Expected value is defined by:

Expected value (EV) = ∑(x \* P(x))

Only the values of x are given, whereas the probabilities are not given.

Therefore, the expected value will be only ∑x

EV = ∑x = 108+110+123+134+135+145+167+187+199

EV = 1308.

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

**Cars speed and distance**



Ans: For Speed, Skewness = -0.895, The value of skewness is negative as the distribution stretches to left more than right. Therefore it is left skewed.

For speed, the kurtosis value= 0.2496, which is positive value so it is called leptokurtic distribution.

For Distance, Skewness =1.2908, The value of skewness is positive as the distribution stretches to right more than left. Therefore it is right skewed.

For distance, the kurtosis value=2.4645, which is also a positive value so it is called leptokurtic distribution.

**SP and Weight(WT)**



Ans: For SP, Skewness = -0.42676, The value of skewness is negative as the distribution stretches to left more than right. Therefore it is left skewed.

For SP, the kurtosis value= -0.86373, which is Negative value so it is called platykurtic distribution.

For WT, Skewness =-1.34755, The value of skewness is negative as the distribution stretches to left more than right. Therefore it is left skewed.

For WT, the kurtosis value=1.152953, which is a positive value so it is called leptokurtic

distribution.

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



Ans: As the distribution of data in the histogram are skewed more to the right of the long tail, It is right-skewed. The mean is larger than the median.



As the minimum and first quartile is near from the median and the maximum and third quartile is far from the median, it is also right skewed. And the data contains many 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: a) 200 as point estimate

94% confidence level

s= 30

n = 2000

Interval Estimate = Point Estimate ± Margin of Error

To find margin of error

We need to find standard error and critical value.

Standard error= SE= s/√(n) = 30/ √(2000)=0.6708

Critical value:

When the data is normally distributed, we will have a symmetrical distribution.

So therefore, 94% will be treated as 97% that means we need to find a z score for 0.97.

From the z score table, Critical value is 1.882

Therefore margin of error= critical value \* standard error = 1.882 \* 0.6708 = 1.2624456

The 94% interval estimate will be 200 ± 1.26244 = [198.737, 201.262]

b) 200 as point estimate

96% confidence level

s= 30

n = 2000

Interval Estimate = Point Estimate ± Margin of Error

To find margin of error

We need to find standard error and critical value.

Standard error= SE= s/√(n) = 30/ √(2000)=0.6708

Critical value:

When the data is normally distributed, we will have a symmetrical distribution.

So therefore, 96% will be treated as 98% that means we need to find a z score for 0.98.

From the z score table, Critical value is 2.05

Therefore margin of error= critical value \* standard error = 2.05 \* 0.6708 = 1.375

Margin error = 1.375

The 96% confidence interval will be 200 ± 1.375 = [198.625, 201.375]

c) 200 as point estimate

98% confidence level

s= 30

n = 2000

Interval Estimate = Point Estimate ± Margin of Error

To find margin of error

We need to find standard error and critical value.

Standard error= SE= s/√(n) = 30/ √(2000)=0.6708

Critical value:

When the data is normally distributed, we will have a symmetrical distribution.

So therefore, 98% will be treated as 99% that means we need to find a z score for 0.99.

From the z score table, Critical value is 2.33

Therefore margin of error= critical value \* standard error = 2.33 \* 0.6708 = 1.5629

The 98% interval estimate will be 200 ± 1.5629 = [198.437, 201.5629].

**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: Mean: 41

Median: 40.5

Mode is not asked but mode=41.

Variance: 25.529

Standard deviation: 5.052664

So as the mean, median and mode are approximately equal, it follows a normal distribution.

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

Ans: Symmetrical distribution.

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

Ans: Positively skewed (Right skewed) distribution.

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

Ans: Negatively skewed (Left skewed) distribution.

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

Ans:Positive kurtosis values indicates that the distribution has heavier tails and a more peaked distribution than the normal distribution, and it is called leptokurtic distribution.

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

Ans: Negative kurtosis values indicates that the distribution has lighter tails and a less peaked distribution than the normal distribution, and it is called Platykurtic distribution.

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



What can we say about the distribution of the data?

Ans: From the distribution of data, we can say that it doesn’t have any outliers.

What is nature of skewness of the data?

Ans: Negative skew(Left skewed).

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

Ans: The IQR of the data will be approximately equal to 8.

Q19) Comment on the below Boxplot visualizations?



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

Ans: When we compare boxplot 1 with box plot 2:

* Same median
* Boxplot 1 is distributed for a lesser range when compared to box plot 2.

Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

Calculate the probability of MPG ofCars for the below cases.

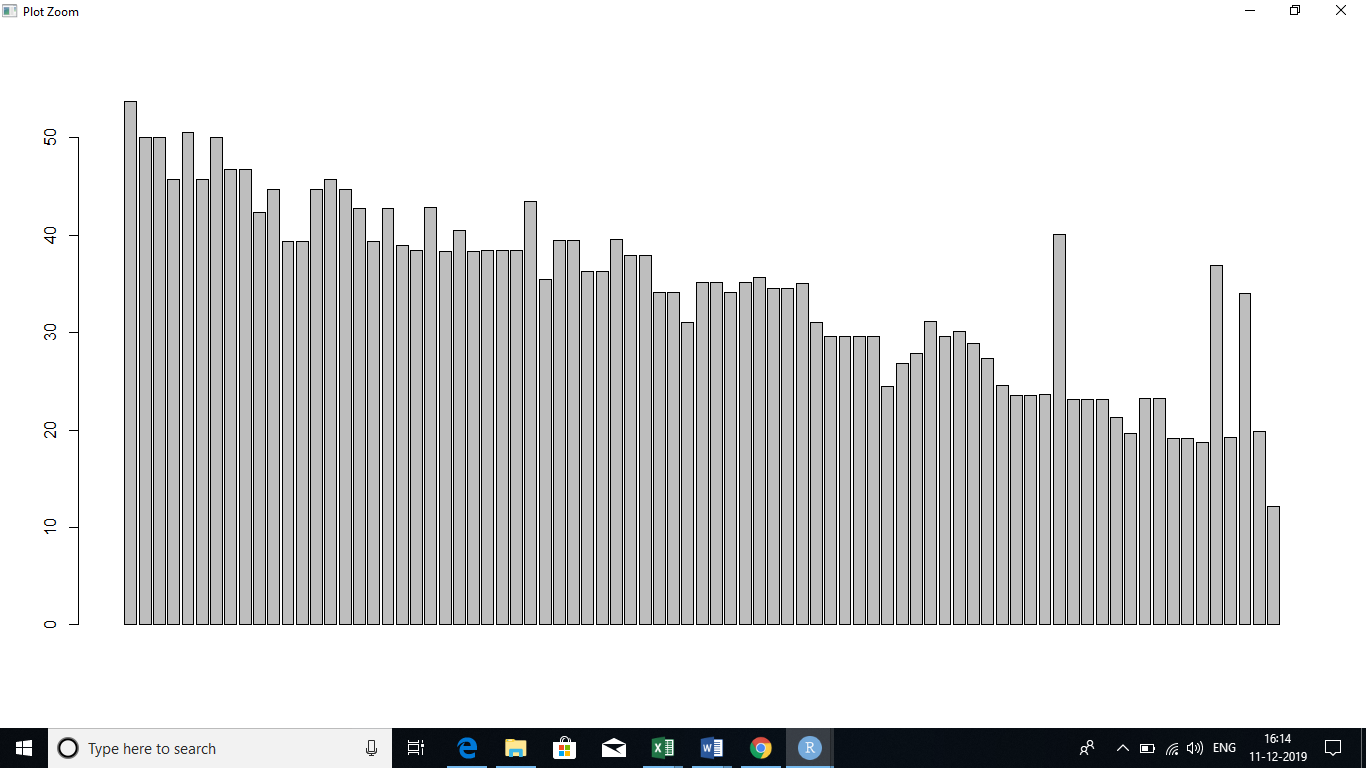
MPG<- Cars$MPG

* 1. P(MPG>38) = 33/81
  2. P(MPG<40) = 61/81
  3. P (20<MPG<50) = 76/81

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv



Ans: The MPG of cars does not follows Normal distribution.

Mean = 34.42208

Median =35.15273

Mode =29.62994

Mean, Median and Mode are not equal. And it does not follow a bell-shaped curve.

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

Dataset: wc-at.csv

Ans: Dataset does not follow normal distribution.

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

Ans: a) 90% confidence interval

When the data is normally distributed, we will have a symmetrical distribution.

So we need to find the z score of 95% that means for 0.95

From z-score table, z-score: 1.644

b) 94% confidence interval

When the data is normally distributed, we will have a symmetrical distribution.

So we need to find the z score of 97% that means for 0.97

From z-score calculator, z-score: 1.88

1. 60% confidence interval

When the data is normally distributed, we will have a symmetrical distribution.

So we need to find the z score of 80% that means for 0.8

From z-score calculator, z-score: 0.84

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

Ans: Sample size, n = 25

Degrees of freedom (df) = n – 1 = 25 – 1 = 24

1. So using t-score table, with 95% confidence interval and degrees of freedom as 24, The t-score will be 2.064
2. With 96% confidence interval and degrees of freedom as 24, The t-score will be 2.1715
3. With 99% confidence interval and degrees of freedom as 24, the t-score will be 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 260days.

Hint: rcode🡪pt(tscore,df)

df🡪 degrees of freedom

Ans: df = 18-1 = 17

µ = 270 days

ẍ = 260 days

σ = 50

t- score: (260-270)/(90/sqrt(18)) =- 0.4714

Cumulative probability = 0.3217

Therefore, if the true bulb were 270 days, there is a 32.17% chance that the average bulb life for 18 randomly selected bulbs would be lesser than or equal to 260 days.